TRANSCRIPTIONAL REGULATION OF GLYCOSYLTRANSFERASE GENES IN MCF-7 HUMAN BREAST CANCER CELL LINE FOLLOWING DRUG TREATMENT

Dissertation presented to the

UNIVERSITY OF CAPE TOWN

In fulfilment of requirements for the degree of

MASTER OF SCIENCE

by

Ju Young Kim

Supervisor: Professor Kevin J. Naidoo

SCIENTIFIC COMPUTING RESEARCH UNIT

Department of Chemistry

UNIVERSITY OF CAPE TOWN

May 2018

DECLARATION

I declare that this dissertation, titled TRANSCRIPTIONAL REGULATION OF GLYCOSYLTRANSFERASE GENES IN MCF-7 HUMAN BREAST CANCER CELL LINE FOLLOWING DRUG TREATMENT, is a presentation of my original research work done at the Scientific Computing Research Unit, Department of Chemistry, University of Cape Town, South Africa. No part of this thesis has been submitted elsewhere for any other degree or qualification. Whenever contribution of others is involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions.

…….

Ju Young Kim

ABSTRACT

Bioinformatics is a subfield in computational science that is principally focused on developing methods and performing data analytics in the areas of proteomics and genomics. In this thesis I draw a link between proteomics and genomics by focusing on the regulation patterns of glycosyltransferase (GT) genes in breast cancer cell line following the treatment with a large set of Food and Drug Administration (FDA) approved drugs. This is based on the understanding that aberrant glycosylation in breast cancer tumours stem from altered GT gene expression.

A major goal of genomic research is the identification of genes that have been differentially expressed under abnormal conditions. A gene expression profile provides a snapshot of the transcriptional level of a cell. A comparative gene expression profile between a diseased and normal state can be used to map out the regulatory mechanisms of disease. In this thesis, the results of Microarray experiments on MCF-7 human breast cancer cell-lines are analysed using statistical and computational tools to identify differentially expressed genes.

Here a bioinformatics analysis of the regulation of GT gene expressions was performed to identify a set of glycosylation related genes with the aim of making an inference about their biological functions. A set of raw gene expression profiles from MCF-7 human breast cancer cell-line treated with different therapeutic drugs were obtained from the Connectivity Map (CMap) database. Initially 7,000 gene expression profiles were used and these were treated by 1,309 different FDA-approved drugs. The number of genes initially was counted up to 22,000. Using the Bioconductor open source software in R statistical programming environment a statistical differential expression analysis followed by several data filtering and pre-processing steps were performed to identify up and down regulated GT genes using. Using non-parametric rank sum meta-analysis three cancer drugs and two non-cancer drugs were identified as effective agents able to control the transcriptional regulatory state of GT genes.

The study concluded by employing co-expression gene module analysis using the Weighted Gene Co-Expression Network Analysis (WGCNA) package on each of the cancer and non-cancer drug treatments. The gene modules discovered from the analysis were used to perform gene ontology enrichment analysis to identify the biological functions where they were significantly enriched in. The co-expression modules where GT genes have been down regulated by the drugs, were

involved in processes such as Wnt signalling and cell surface pattern recognition receptor signalling important for cancer development. Immune response and apoptotic processes in the cell were identified from co-expression modules where GT genes were up regulated. This key finding that the GT gene expressions are markers for treatment analysis points to their use in drug development studies. The second more direct finding is that non-breast cancer specific FDA-approved drugs may have a role in treating breast cancer and may be the subject of future drug repurposing strategies.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest appreciation to my supervisor, Professor Kevin J. Naidoo for his continuous support, guidance and advice with my research and thesis, and for his patience, motivation, and immense knowledge. Without his help and guidance, this would not have been possible for me.

Besides my advisor, I am also thankful to Dr. Christopher Barnett for his helps in many ways in the work.

My sincere thanks also go to Dr. Christos Ferles for sharing many great insights and having interesting conversations throughout my research work.

I would also like to thank all my friends and colleagues from the Scientific Computing Research Unit (SCRU) in many ways for all the help and supports. I also thank Louise Bezuidenhout, Lisl George and Lydia Dreyer at SCRU and Deidre Brooks and all the wonderful people at the chemistry department for taking care of all the administrative duties and for being immensely helpful and friendly.

I thank the Scientific Computing Research Unit and the University of Cape Town for the opportunity, facilities, environment and the National Research Foundation and the South African Research Chairs Initiative (SARChI) for my master's scholarship.

A huge thank goes to my family for standing by me and providing all the supports they could have provided for me during my last two years of research work.

Last but not the least, I would like to praise and thank our God for giving me the strength, wisdom, ability, confidence and opportunity to undertake this research study and complete it satisfactorily. Without his blessings, this achievement would not have been possible.

ABBREVIATIONS

BH	Benjamini-Hochberg
CAZy	Carbohydrate-active enzymes
CMap	Connectivity map
ComBat	Combating batch effects
DCM	Differentially co-expressed modules
DEG	Differentially expressed genes
DNA	Deoxyribonucleic acid
ER	Endoplasmic reticulum
FDR	False discovery rate
FUT	Fucosyltransferase
GalNAc	N-Acetylglucosamine
GEO	Gene Expression Omnibus
GlcNAc	N-Acetylglucosamine
GO	Gene ontology
GT	Glycosyltransferase
KEGG	Kyoto encyclopedia of genes and genomes
LIMMA	Linear models for microarray
MCF-7	Michigan cancer foundation-7
MM	Mismatch
NGS	Next-generation sequencing
PCA	Principal component analysis
PDB	Protein data bank
PM	Perfect match
PTM	Posttranslational modification
RMA	Robust multi average
RNA	Ribonucleic acid
RTK	Receptor tyrosine kinase
SQL	Structured query language
ST	Sialyltransferase
TCGA	The Cancer Genome Atlas
WGCNA	Weighted gene co-expression network analysis

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 COMPUTATIONAL SCIENCE AND BIOINFORMATICS

1.1 Introduction

Computational Science by its definition is a multidisciplinary field that incorporates advanced computing power and numerical analytics capabilities to solve complex scientific problems. The recent advancement in computing power and development of methodologies have radically changed the way how we collect complex high-dimensional data. Large data sets collected using advanced computing power have benefited research across all computational science that include computational biology where the focus is the development of models for molecular biology and the use of data analytics for genomics. Large public databases have been created to store high-dimensional data often making these publicly available. The availability of large databases enabled the participation of a community of researchers, that develop tools for analysis and explore the data sets so making meaningful scientific contribution unencumbered by outdated traditional tools of a given field. (Peng, 2011)

The advancement of computational science tools has contributed immensely to the development of modern biology. Many aspects of biological research incorporate mathematical, statistical models that have been derived from and computational methods that store, exploit, analyse massive amount of biological data for the purpose of answering fundamental biological questions. Regardless of whether the problem is simple or complex, the computational tools and mathematical concepts have become integral to modern biological research.

Bioinformatics is an interdisciplinary field that implements statistical and computer science methods to biology. Bioinformatics recently gained its popularity because of the unprecedented amounts of biological data generated from advanced genomic methods, a rapid increase in computing powers and the development of analytics tools. Developing and using computational tools in molecular biology requires knowledge from a wide range of disciplines including mathematics, statistics, and information technology. This overlap of the computer and mathematical science tools applied to the domain science of molecular biology came to be known as bioinformatics. Computers are used to store, retrieve, manipulate, and distribute macromolecular information on biomolecules DNA, RNA, proteins and now more recently

glycans (Barnett, Aoki-Kinoshita, & Naidoo, 2016). Because genomes are made from simple four letter alphabets. In the field of genomics data analysis, the computational routines traditionally used for text analysis can be transplanted. Complex methods can be assembled into bioinformatics analytic pipelines, making data mining in genomics easily accessible to mathematical novices. (Xiong, 2006) (Luscombe, Greenbaum, & Gerstein, 2001) Figure 1.1 describes definition of computational science.

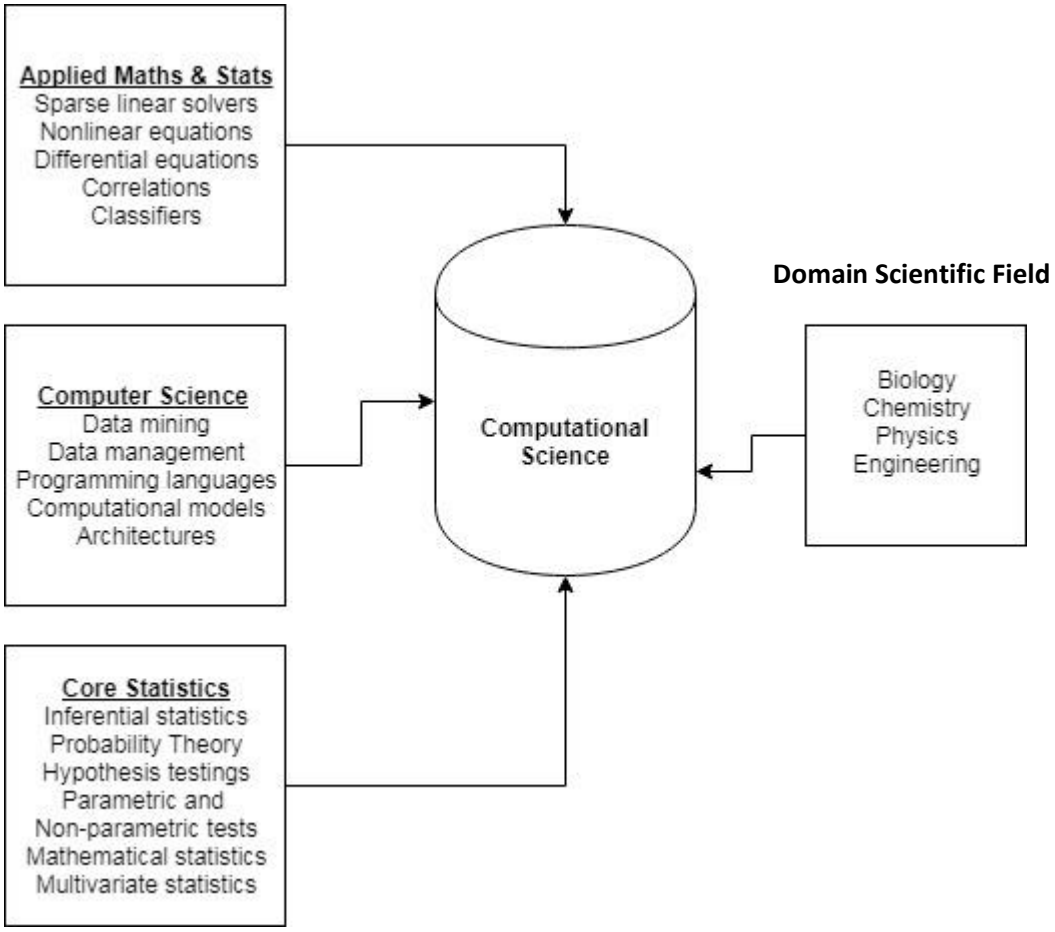


Figure 1.1 Components of computational science that incorporates mathematics, statistics, computer science and domain scientific fields. When the focus of the domain scientific field is uniquely from molecular biology and genomics, the combination is called bioinformatics.

Bioinformatics is often referred to as computational biology, but the two fields are different in terms of biological problems dealt with. In bioinformatics, the topics focused are usually in sequence, structural, and functional analysis of genes and genomes and the corresponding product

known as proteins. Computational biology covers all biological areas that involve computation, for instance, population dynamics, phylogenetic, behavioural studies etc. (Xiong, 2006) Although the two terms have distinct characteristics, there are still many significant overlaps and activities at their interface. (Rao, Das, Rao, & Srinubabu, 2008) Ultimately the goals of bioinformatics are to utilize new technologies to better understand cells and their functions at a molecular level. (Xiong, 2006) From functional genomics to molecular sequence analysis and structural bioinformatics, this new and exciting sub-field of computational science have several aims. First, to organize and store biological data which can be accessed by other researchers in the field to conduct experiments following their unique protocols and submit new entries. Second, to develop efficient tools that will aid researche with their data mining and analysis activities. Third, use the available data to perform high-throughput analysis and biologically interpret meaningful results. Fourth, to aid the researchers in pharmaceutical industries to design and develop novel drugs for deadly diseases. (Rao et al., 2008) Before a deeper scope of bioinformatics can be explained, the next few paragraphs will cover topics including, big data in biology and recent next-generation sequencing techniques. Also, critical role of bioinformatics in the current setup of clinical research will be discussed.

1.2 Big data in biology

With the advent of high-throughput genomics and technologies, biologists have joined the big data club. (Marx, 2013) The new technologies allow biologists to generate enormous amounts of data in a cost-effective manner and this low-cost data generated, led to a big data era in biology. The data generated range from genomic sequence to images of physiological structures and in between the two types of data, factors such as mRNA expression, transcription factor binding, protein expression are measured as explained by the central dogma of molecular biology. (Gesing, Connor, & Taylor, 2015) Following on Human Genome Project that took 15 years and over $2 billion to complete, the Genome Analyzer was launched by Solexa, a biotechnology firm (now Illumina) in 2006. Next-generation instruments such as Genome Analyzer has revolutionized the genomics field in that, it could do the same task of sequencing the genome in less than 3 days for less than $1000 (Gesing et al., 2015) (Marx, 2013).

The next-generation sequencing (NGS) technologies have created various biological and clinical datasets. The Cancer Genome Atlas (TCGA) is a project initiated in 2005, supervised by National Cancer Institute's Center for Cancer Genomics and the National Human Genome Research Institute (NHGRI) that contains publicly available dataset of somatic mutations, copy number variation, mRNA expression, protein expression, and histology slides for approximately 7,000 human tumours (Gesing et al., 2015) (Tomczak, Czerwińska, & Wiznerowicz, 2015). Encyclopaedia of DNA elements (ENCODE) is a public research project launched by US NHGRI in 2003 that has generated more than 2,600 genomic datasets from sequencing methods such as ChIP-Seq, RNA-Seq, ChIA-PET and CAGE. With a quantitative increase in genomic data, there were needs for repositories to store the data to analyse them. The European Bioinformatics Institute (EBI) offers repositories that store biological data including genomic data. There are many publicly available databases that contain genomic data produced in the labs of individual researchers and the data was made available for reproducible research, when their publications were released. (Gesing et al., 2015) For instance, Gene Expression Omnibus (GEO) is a public repository that archives high throughput data such as microarray data and Array Express, is a public repository that contains 1.3 million gene expression data from more than 45,000 laboratory experiments. (Rustici et al., 2012) The NGS platforms have enabled the creation of such public databases. The next few paragraphs will discuss some aspects of NGS and its challenges.

1.2.1 Next-generation sequencing

The "first generation sequencing" techniques were initiated from Sanger's sequencing method. (Sanger, Nicklen, & Coulson, 1977) The success of human genome project has led the researchers to identify genome sequences in other species as well and several years later, the next-generation sequencing (NGS) method, called "second generation sequencing" techniques have appeared that revolutionized biological research fields. (Metzker, 2010) The next-generation sequencing methods performed sequencing tasks much quicker and cheaper than conventional sequencing techniques. With these advantages of the methods, the NGS technologies have been widely used for many applications including, transcriptome profiling that sequences cDNA to get a sample's RNA content, which has become the standard method for measuring RNA expression level (Xuan, Yu, Qing, Guo, & Shi, 2013) (Schuster, 2007) (Zhao, Fung-Leung, Bittner, Ngo, & Liu, 2014).

There are numerous NGS platforms to power biological research, including Roche/454 sequencing system, Illumina/Solexa and Life Technologies/SOLiD (Schuster, 2007) (Dillies et al., 2013). The technical details of these major NGS platforms can be found in (Metzker, 2010). Each platform has its own advantages and disadvantages and the performance of each sequencing platform is described in table 1.1 along with their pros and cons.

Table 1.1 Overview of performance of different NGS platforms[a)].

Company	Platform	Amplification	Sequencing	Read Length	Time per run	Overall error rate	Pros	Cons
Roche 454	GS FLX Titanium XL	Emulsion PCR	Pyrosequencing	Up to 1kb	700 Mb/23 h	0.5%	Long read	High cost/MB
Illumina	Genome Analyzer	Bridge PCR	Sequencing by synthesis with reversible terminator	35 – 150 bp	10-95 Gb/2-14 days	0.2%	Most reads, GB/day, low cost/MB	High capital cost and computation needs
Life Technologies/Applied Biosystems	5500xl SOLiD™ system	Emulsion PCR	Sequencing by ligation	35 – 75 bp	10-15 Gb/day	0.1%	High accuracy	Short reads, more gaps in assemblies
Life Technologies/Ion Torrent	Ion Proton™ sequencer	Emulsion PCR	Ion semiconductor sequencing	Up to 200 bp	Up to 10 Gb/2-4 h	1%	Short time, low cost per sample	Long time of sample preparation

a) information is based on company sources and www.molecularecologist.com.

The platforms presented in table 1.1 have error types such as substitution, indel, A-T bias and deletion with their respective error rates. These errors must be carefully assessed and corrected so the potential impact on downstream analysis can be minimized. (Xuan et al., 2013)

The Illumina's platforms have been widely used for diverse experimental purposes because of large amount of read production per price, which is mostly desired feature for NGS platform. Analysing and interpreting NGS data accurately and efficiently is crucial to apply NGS technologies in personalized medicine and bioinformatics provides tools and analytical capabilities to translate the data for personalized clinical medicine. (Hong et al., 2013)

1.2.2 Computational challenges

A significant advance in biological research has been the large volumes of data produced at historically short experimental times forms NGS technologies. Analysing these volumes of data appeared at first as a computational challenge. NGS with its deep sequencing high throughput methods, generate hundreds of millions of short sequence reads (Xuan et al., 2013) requires data storing, transferring and analyses. The Illumina sequencing platform for example generates 400 million reads (100 bp paired end reads) in one lane that is equivalent to 40 GB of data. A typical sequence data analysis utilizes a few hundred GBs of disk space, memory and requires processing on multi-CPU computer nodes. For instance in de novo assembly, AbySS genome assembly tool requires at least 150 GB disk space and 24 cores for sequence data obtained from a single lane whereas, Trinity genome assembly requires at least 200 GB of memory. The gene expression analysis using RNA-Seq for example, requires computer clusters with nodes that have several hundred GB of memory, 2x12core CPUs and Tb disk space making this a costly supercomputing environment for many small laboratories. (Liu, 2017) Thus, prior to making decision regarding which NGS platform to use, it is critical to consider data storage and memory requirements. (Hong et al., 2013)

1.3 Trends in bioinformatics

Since the completion of the human genome project, the paradigm has shifted to a post-genomic era with an availability of large scale genomic data for biomedical informatics research. (Knaup et al., 2004) The National Center for Biotechnology Information (NCBI) especially, has contributed greatly in maintaining and growing collection of databases of genetic sequence and protein data along with other bioinformatics databases and this has transformed a conventional

genomic research into a computational domain which has become indispensable in life science research. (Knaup et al., 2004) With this, bioinformatics currently has roles in numerous field of life sciences research.

Drug discovery has largely benefited from bioinformatics approach. High-throughput experiments have characterized drug target and disease process at a molecular level. The availability of massive screening libraries and drug databases have led to an efficient lead identification stage of drug development. Computational reconstruction and characterization of molecular pathways of different diseases using high-throughput experiments have been promising in drug development. (Fischer, 2005) Using computational approaches, pharmaceutical companies were able to identify novel drug target, disease state and different pathways related to the disease. The bioinformatics approach played a key role in both lead drug identification and drug repurposing in pharmaceutical industries. (Knaup et al., 2004) As the field of disease genetics and pharmacogenetics evolve, with bioinformatics approach, it is anticipated to enhance the scientific understanding of relationship between disease susceptibility and drug response for developing a novel target therapy. (Rao et al., 2008)

1.3.1 Bioinformatics in clinical research

The bioinformatics approach plays a significant role in identifying molecular pathways and mechanisms that are important in clinical research. Of particular relevance here are studies that focus on the role of carbohydrates in cancer. Recently, Dong et al (2017) have used powerful bioinformatics pipelines combined with immunohistochemistry (IHC) analysis to identify key genes that affect carbohydrate macromolecular structures implicated in the development of gastric cancer. In the post-genomic era, much attention has turned to post-translational modifications (PTMs) and the most common by far is glycosylation. (Kikuchi & Narimatsu, 2006) Of greater consequence is the knowledge that aberrant glycosylation in tumours (due to altered glycosyltransferase (GT) gene expression) plays an important role in the development of cancer. (Ashkani & Naidoo, 2016)

(Dong et al., 2017) used genetic data for gastric cancer samples and healthy normal samples that are publicly available on the TCGA database (section 1.2) to perform differential gene expression analysis between the two samples with the aim of identifying GT genes that have been up and

down regulated. The expression pattern of the GT genes was examined to identify GT genes that could act as potential biomarkers. Their analysis revealed that Protein O-fucosyltransferase-1 (POFUT1), that catalyses the addition of O-fucose to epidermal growth factor-like (EGF) repeats in endoplasmic reticulum, is over-expressed. The gene set enrichment analysis identified that POFUT1 gene was involved in cell cycles and cell carcinoma functions in gastric cancer. (Dong et al., 2017) Through immunohistochemical analysis, they verified that POFUT1 gene was identified in various tumour samples. It was expected from the study that knockdown of POFUT1 gene could inhibit progression of gastric cancer and with this, they have concluded that this key GT gene could potentially be a biomarker for gastric cancer, though further investigations are needed.

(Ashkani & Naidoo, 2016) have examined the expression profiles of GT genes to classify cancer types. In a pioneering use of TCGA database to analyse the relative expression profiles of 210 GT genes from 1,893 cancer samples they discovered that they were able to classify breast, ovarian, glioblastoma, kidney, colon and lung cancers as well as discover subtypes for them. The subtype discovery allows prognosis based on the role of glycosylation in different cancer types making possible a personalised treatment strategy for cancer patients.

The few examples described above illustrate that bioinformatics methods provide a way to investigate the role of postgenomic events that could significantly contribute to a research into personalised treatment.

1.4 Bioinformatics applications and methodologies

There are two main foci in bioinformatics research. The first is the development of software, computational tools and databases for biological data. The second revolves around the applications of these tools and databases developed to gain biological insight not possible from singular experimental observations. The software development is based on implementing sophisticated statistical models and computational algorithms that can be used to create stable databases and perform accurate and fast analyses of biological data. The tools created are developed for application in three major research areas namely, sequence analysis, functional analysis and structural analysis. Often these three applications are combined to produce an integrated result.

For instance, the prediction of protein structure needs the results from sequence alignment analysis and cellular function prediction requires methods from all three areas. (Sensen, 2008) (Xiong, 2006) Figure 1.2 provides an overview of bioinformatics application ecosystem.

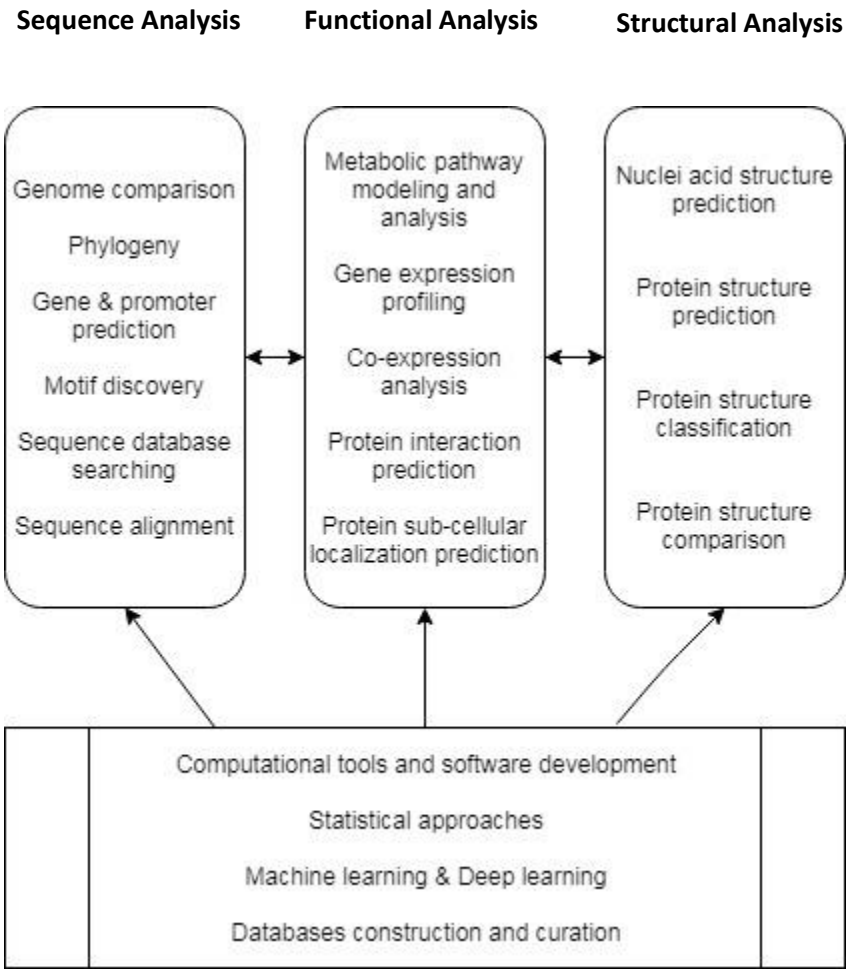


Figure 1.2 Overview of an ecosystem of bioinformatics fields. Bioinformatics tools development becomes the foundational basis for other analyses. Three applications are: sequence, functional, structural analysis and the three applications have intrinsic connections between them.

1.4.1 Sequence analysis

Sequence analysis deals with DNA, RNA or peptides (protein) sequences to understand its features, function and structures. There are many applications of sequence analysis, for instance, one can identify sequence differences and variations such as single nucleotide polymorphism (SNP) to

acquire a genetic marker or one can compare different sequences to identify homologous sequence that displays sequence similarity. (Sensen, 2008) (Xiong, 2006) (Kapetanovic, Rosenfeld, & Izmirlian, 2004) Some of the popular methods for sequence comparison includes, local and global sequence alignments (figure 1.3).

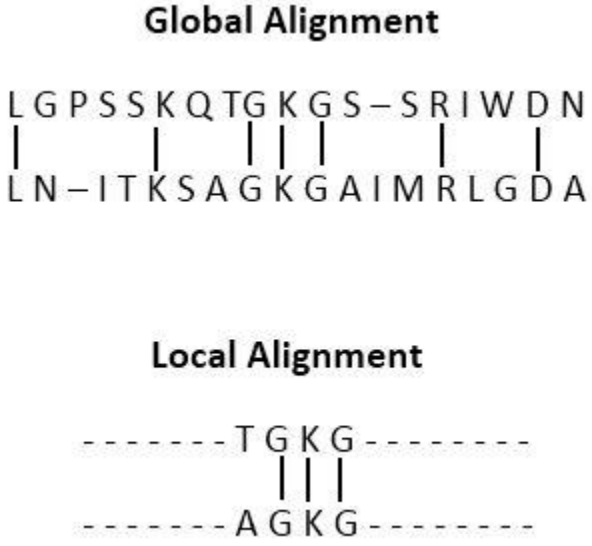


Figure 1.3 An example of pairwise sequence comparison methods. The global sequence alignment method looks at all the residues. The local alignment includes portions of residues of the two sequences that have the highest similarity. The pipe character '|' indicates an identical residue.

1.4.2 Functional analysis

Areas of functional analysis includes, gene expression profiling, protein-protein interaction prediction, metabolic and functional pathway analysis etc. As an example, functionally important genes in a disease can be analysed using computational methods and statistical approaches to identify how many genes are expressed for a specific function. Once the expression patterns of the genes are identified, it is possible to select a few genes that have shown significant amount of regulation and the set of genes can then undergo enrichment and pathway analysis to associate them to specific biological functions in a cell. The results of gene regulation can be visualized in a heatmap (figure 1.4). (Parmigiani, Garrett, Irizarry, & Zeger, 2003) (Ignatenko et al., 2009)

a) **b)**

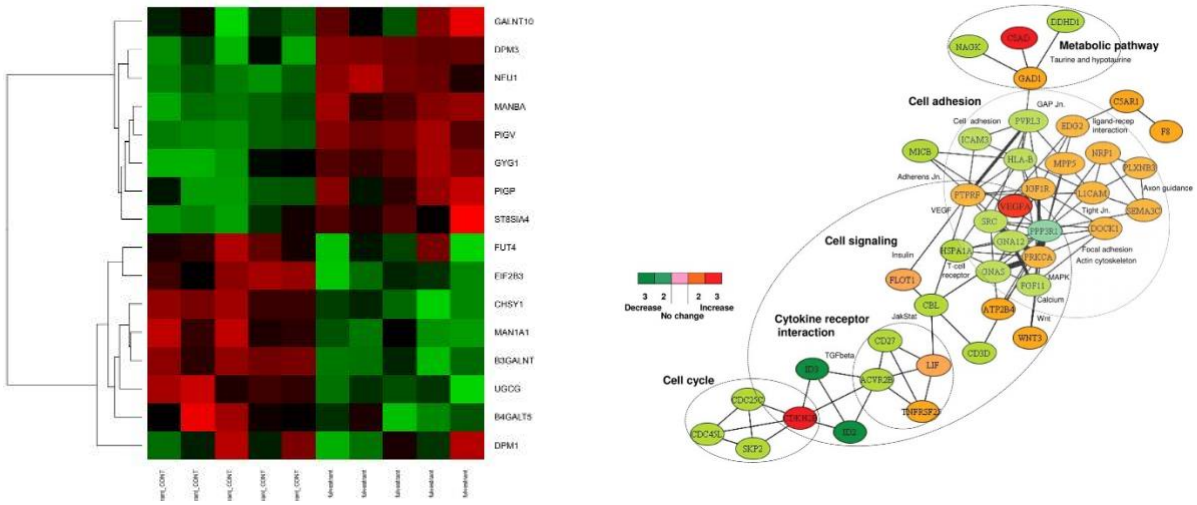


Figure 1.4 An example of methods in gene expression analysis. a) Gene expression patterns are visualized under the heatmap to identify up and down regulated genes. b) The identified up and down regulated genes from the heatmap undergo pathway and functional enrichment network analysis to associate the group of genes to a specific pathway and biological functions in a cell (Massimino, 2017).

1.4.3 Structural analysis

Finally, structural analysis includes protein and nucleic acid structure modelling and analysis, comparison, classification and prediction. Studies involving protein structure analysis are often collectively referred to as proteomics. Proteomics methods can be used to identify therapeutic targets, the development of biomarkers and an examination of protein alterations that is not possible to gauge from DNA translation. (Baxevanis & Ouellette, 2004) (Goeta, 2002) For instance, Protein Data Bank (PDB) is a popular crystallographic database for three-dimensional structural data of large biomolecules and various protein structures can be obtained. The protein structures obtained can be modelled by running different simulations of docking to identify its active sites and develop new therapeutic targets using computational tools and sophisticated bioinformatics software.

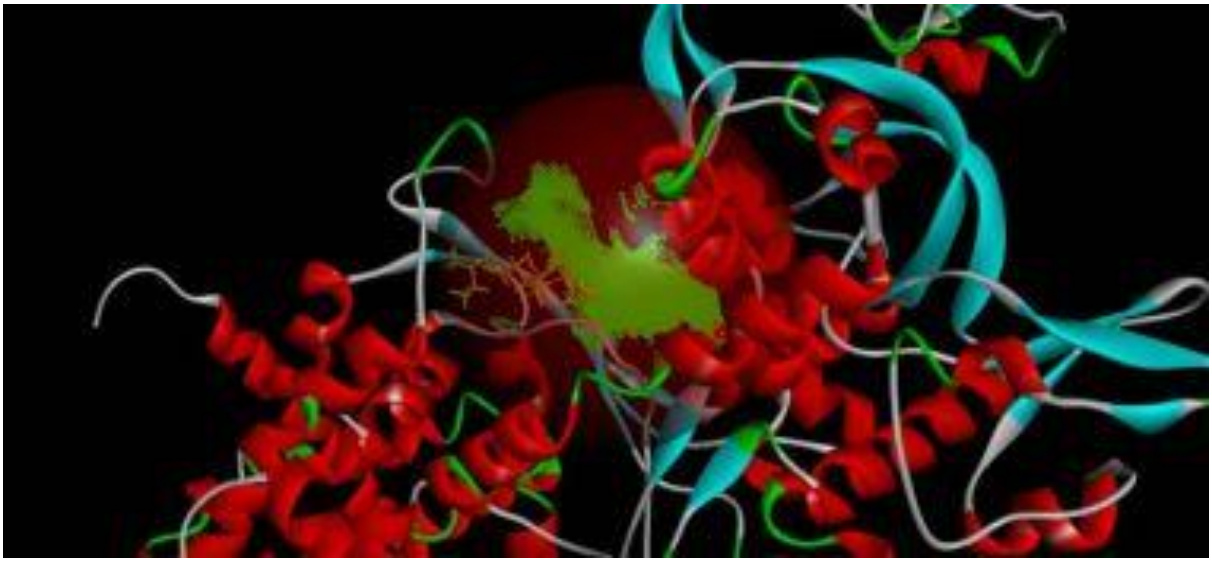


Figure 1.5 An example of molecular docking of batumin antibiotics to staphylococcus aureus active site using Discovery Studio program developed from Accelrys. The program offers functionalities for molecular docking, protein-binding site properties and complex ab initio simulations and many more (Tikhvinskiy & Porozov, 2013).

1.4.4 Biological databases

Following the Human Genome Project, the introduction of Next-Generation Sequencing (NGS) methods led to an exponential growth in, genomic data. New efficient computational methods were needed to store and process the NGS generated data making a wide range of databases available to biologists to analyse and make discoveries from. The computerized archives offer a large quantity of structured and standardized data that are easily retrieved. (Goeta, 2002) (Xiong, 2006) Databases can be used to identify new connections for instance, raw sequences from a database can be compared using computational methods to identify homologous motifs or gene expression patterns of different samples can be compared to discover mechanisms for diseases. This type of database application is aptly named *knowledge discovery*. (Isewon, 2009) Currently a relational database is the most popular model. Typically, these can be created using a structured query language (SQL). (Sensen, 2008) Here a set of tables is used to organize data instead of using a single table with flat files. The relational database model offers great querying performance for large data sizes. Further, when different biological databases are interconnected, the entries from different databases can relate to each other which opens up another avenue to knowledge discovery

using information retrieval. The complexity of biomedical data is in the use of graph database as a replacement for relational database, such as platforms like Neo4J.

1.4.5 Software development and statistical approaches

The bioinformatician's tasks includes, data storing, management and modelling. While the biological databases have been created (section 1.1.4) to address the data storing and management tasks, it is crucial to use efficient modelling methods to effectively analyse the vast amount of biological data. To accomplish this, statistical and computational methods must be incorporated to develop useful statistical and mathematical models embedded in bioinformatics software that facilitate data analytics. There are a number of programs developed for specific applications in bioinformatics. For instance, the Bioconductor project, which will be discussed in detail in chapter 2, is an open-source software package developed to perform various genomic data analysis. The Bioconductor requires an R statistical programming environment making many useful statistical modelling techniques possible. (Gentleman et al., 2004) Machine Learning (ML) techniques and innovative deep learning methods have recently gained popularity amongst bioinformaticians. For instance, microarray data gene expression pattern identification, genetic network induction, classifier development for cancer subtype prediction and clustering of genes all use machine learning methods and the results are promising. (Ashkani & Naidoo, 2016) (Larranaga et al., 2006) (Min, Lee, & Yoon, 2017) Overall, bioinformatics software development using tailored computational algorithms and rigorous statistical models transforms traditional analytical pipelines to efficiently accelerate biological data modelling and analysis.

1.5 Limitations and issues in bioinformatics

Although an informatics approach to analysing genomic data provides powerful pathways to discovery but it must be cautioned that an over-reliance on computational outputs without a clear understanding of the statistical methods underlying the techniques and the quality of data being analysed is foolhardy. A case in point is that poor quality data generated by NGS techniques can lead to a catastrophic misinterpretations. (Dillies et al., 2013) Before embarking on an extensive computational experiment, the raw data generated from NGS experiments must be evaluated to

uncover errors due to bias and/or noisy data. If these are not corrected or if erroneous raw data are used in the downstream analysis, the entire biological interpretation will be faulty. The computational algorithms are not always efficient and they could lead to false predictions. (Sensen, 2008) (Baxevanis & Ouellette, 2004) (Xiong, 2006) Often sophisticated computational algorithms that generate accurate results are slow as they rely on the availability of large computing resources. Comparatively, faster algorithms are found to generate less accurate results. The novice practitioner may be faced with the dilemma faced with a choice of speed vs. accuracy. To overcome this pitfall, it is often recommended to perform the analysis using multiple pipelines, algorithms or programs. Multiple evaluations are then possible as the results generated from different methods can be compared. (Goeta, 2002)

1.6 Aims and objectives

The aim of this research was to identify and examine gene expression regulation of key glycosyltransferase (GT) genes and their associated biological functions with other genes from the MCF-7 human breast cancer cell line treated with therapeutic Food and Drug Administration (FDA) approved drugs. This required employing functional analysis methodologies (section 1.4.2). Moreover, a specific objective was to understand the role of carbohydrate related genes in breast cancer. This was done by investigating the effect of known drugs administered to cancer cell lines on the regulation of glycosyltransferase genes.

The first methodological objective was to pre-process raw gene expression data retrieved from the Connectivity Map (CMap) database using standard microarray data analytics pipeline to generate a suitable expression matrix for statistical analysis.

The second methodological objective was to perform differential gene expression analysis using LIMMA tool from Bioconductor project in R statistical computing environment and filter out GT genes for subsequent analysis.

Following the methods employed above the third project objective was to select therapeutic drugs that contributes the most to the regulation of GT genes by performing meta-analysis that incorporates two-groups Wilcoxon rank sum test. Rank sum test is used to statistically test if the drug had significantly affected GT genes in comparison to non-GT genes.

Finally, the aim was to identify co-expression modules of gene expression data from selected drugs using the WGCNA clustering method and associate the co-expression modules to biological functions using a gene ontology enrichment analysis. Differentially expressed GT genes within the co-expression module were identified. The GT genes regulatory pattern and the involvement of GT genes in the co-expression modules with its associated biological function were used to gain insight into how carbohydrate related genes collaborate with other oncogenes to contribute to breast cancer progression. Further it was hoped to provide possible insight into the discovery of presently approved drugs that could be repurposed for breast cancer treatment.

1.7 Thesis synopsis

In Chapter 2, I discuss relevant bioinformatics gene expression methods, and how to utilize high-throughput technologies such as microarray to examine and quantify gene expression of a sample. The biology underpinning a gene expression is briefly described. The bioinformatics workflow for microarray data analysis such as pre-processing, statistical methods, analytics environment and class discovery topics and methods such as WGCNA can be used to filter gene expression data.

In Chapter 3, I discuss overview of the glycobiology of breast cancer. Molecular mechanisms in breast cancer and its treatment in the context of abnormal glycosylation in cancer is discussed. The role of carbohydrates, cellular function and regulation of glycosyltransferase (GT) genes are summarised as well. Literatures regarding the involvement of some of the most well-known GT genes in breast cancer is described. GT genes have roles in abnormal glycosylation which results in breast cancer and their functions and contributions to breast cancer will be analysed.

In Chapter 4, I discuss how gene expression data for this research work was prepared using the techniques described in chapter 2. An overview of the Connectivity Map (CMap) is given and its database is explained along with research uses of CMap data. CMap Data cleaning and preparation will be done. A detailed information on therapeutic drugs used to treat the breast cancer cell line will be discussed. Several data pre-processing steps will be performed including, data quality checking, normalization, batch effect with principal component analysis and gene annotations. The prepared gene expression data can be used to perform further differential expression analysis.

Hence, the generation of suitable gene expression matrix for statistical analysis was the main goal of this chapter.

In Chapter 5, I discuss statistical analyses of differentially expressed genes (DEGs) from the expression matrix generated in chapter 4. Several statistical analysis tools for gene expression such as LIMMA are described. LIMMA tool will be used to identify statistically significant differentially expressed GT genes and drugs that have significant effect on the regulation of GT genes will be selected. The resulting drug treated samples will further be analysed using WGCNA method to identify co-expression modules of genes from the treatment. Finally, Gene Ontology (GO) enrichment analysis will be performed to identify biological functions associated with co-expression modules. The GT genes involved in each of the co-expression module will be identified and relate them to the biological functions for interpretation. In this way, the roles of GT genes in breast cancer and insight into how to design a novel drug can be obtained.

1.8 References

1. Ashkani, J., & Naidoo, K. J. (2016). Glycosyltransferase gene expression profiles classify cancer types and propose prognostic subtypes. *Scientific reports, 6*, 26451.
2. Barnett, C. B., Aoki-Kinoshita, K. F., & Naidoo, K. J. (2016). The Glycome Analytics Platform: an integrative framework for glycobioinformatics. *Bioinformatics, 32*(19), 3005-3011.
3. Baxevanis, A. D., & Ouellette, B. F. (2004). *Bioinformatics: a practical guide to the analysis of genes and proteins* (Vol. 43): John Wiley & Sons.
4. Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., . . . Estelle, J. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics, 14*(6), 671-683.
5. Dong, S., Wang, Z., Huang, B., Zhang, J., Ge, Y., Fan, Q., & Wang, Z. (2017). Bioinformatics insight into glycosyltransferase gene expression in gastric cancer: POFUT1 is a potential biomarker. *Biochemical and biophysical research communications, 483*(1), 171-177.
6. Fischer, H. P. (2005). Towards quantitative biology: integration of biological information to elucidate disease pathways and to guide drug discovery. *Biotechnology annual review, 11*, 1-68.
7. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., . . . Gentry, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology, 5*(10), R80.

8. Gesing, S., Connor, T. R., & Taylor, I. (2015). Genomics and Biological Big Data: Facing Current and Future Challenges around Data and Software Sharing and Reproducibility. *arXiv preprint arXiv:1511.02689*.
9. Goeta, B. (2002). Bioinformatics-Sequence and Genome Analysis: Henry Stewart Publications.
10. Hong, H., Zhang, W., Shen, J., Su, Z., Ning, B., Han, T., . . . Tong, W. (2013). Critical role of bioinformatics in translating huge amounts of next-generation sequencing data into personalized medicine. *Science China Life Sciences, 56*(2), 110-118.
11. Ignatenko, N. A., Yerushalmi, H. F., Pandey, R., Kachel, K. L., Stringer, D. E., Marton, L. J., & Gerner, E. W. (2009). Gene expression analysis of HCT116 colon tumor-derived cells treated with the polyamine analog PG-11047. *Cancer Genomics-Proteomics, 6*(3), 161-175.
12. Isewon, I. (2009). *DESIGN AND DEVELOPMENT OF THE AFRICAN PLASMODIUM FALCIPARUM DATABASE–(afriPFdb)*. Covenant University.
13. Kapetanovic, I. M., Rosenfeld, S., & Izmirlian, G. (2004). Overview of commonly used bioinformatics methods and their applications. *Annals of the New York Academy of Sciences, 1020*(1), 10-21.
14. Kikuchi, N., & Narimatsu, H. (2006). Bioinformatics for comprehensive finding and analysis of glycosyltransferases. *Biochimica et Biophysica Acta (BBA)-General Subjects, 1760*(4), 578-583.
15. Knaup, P., Ammenwerth, E., Brandner, R., Brigl, B., Fischer, G., Garde, S., . . . Singer, R. (2004). Towards clinical bioinformatics: advancing genomic medicine with informatics methods and tools. *Methods Archive, 43*(3), 302-307.
16. Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., . . . Pérez, A. (2006). Machine learning in bioinformatics. *Briefings in bioinformatics*, 86-112.
17. Liu, Z. J. (2017). *Bioinformatics in Aquaculture: Principles and Methods*: John Wiley & Sons.
18. Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods of information in medicine, 40*(4), 346-358.
19. Marx, V. (2013). Biology: The big challenges of big data: Nature Publishing Group.
20. Massimino, L. (2017). In silico gene expression profiling in Cannabis sativa. *F1000Research, 6*.
21. Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics, 11*(1), 31.
22. Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. *Briefings in bioinformatics, 18*(5), 851-869.
23. Parmigiani, G., Garrett, E. S., Irizarry, R. A., & Zeger, S. L. (2003). The analysis of gene expression data: an overview of methods and software *The analysis of gene expression data* (pp. 1-45): Springer.

24. Peng, R. D. (2011). Reproducible research in computational science. *Science, 334*(6060), 1226-1227.
25. Rao, V. S., Das, S. K., Rao, V. J., & Srinubabu, G. (2008). Recent developments in life sciences research: Role of bioinformatics. *African Journal of Biotechnology, 7*(5).
26. Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., . . . Keays, M. (2012). ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic acids research, 41*(D1), D987-D990.
27. Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences, 74*(12), 5463-5467.
28. Schuster, S. C. (2007). Next-generation sequencing transforms today's biology. *Nature methods, 5*(1), 16.
29. Sensen, C. W. (2008). *Essentials of genomics and bioinformatics*: John Wiley & Sons.
30. Tikhvinskiy, D. A., & Porozov, Y. B. (2013). Bioinformatics and tools for computer analysis and visualization of macromolecules. *Russian Open Medical Journal, 2*(1), 0101-0101.
31. Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology, 19*(1A), A68.
32. Xiong, J. (2006). *Essential bioinformatics*: Cambridge University Press.
33. Xuan, J., Yu, Y., Qing, T., Guo, L., & Shi, L. (2013). Next-generation sequencing in the clinic: promises and challenges. *Cancer letters, 340*(2), 284-295.
34. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS one, 9*(1), e78644.

2 GENE EXPRESSION ANALYSIS TECHNIQUES

2.1 Gene expression analysis overview

The genomics research aims to study the genes, mutation, polymorphisms, and phylogenetic relationship of an organism from its original DNA sequences. One important area of genomics research is the analysis of transcriptional regulation of genes known as 'gene expression'. An examination of gene expression pattern of genes, in response to number of clinical conditions, natural or toxic agents and biological processes at different times, can identify what genes are up and down regulated by the condition. Based on the up and down regulated gene patterns, the key genes involved in cellular functions can be identified. (Suárez, Burguete, & McLachlan, 2009) For instance, the key genes identified can be used to develop a novel biomarker, or they can be used as a specific drug target in certain diseases. The application of gene expression is countless and the insights it can provide are enormous. For this reason, a high-throughput microarray technology that simultaneously measures the expression levels of thousands of genes has been developed.

Recently, a more powerful RNA-Seq technology was introduced in the biotech market, but the microarray technology has still shown promising results in the field of genomics research, particularly in the cancer research and it is expected that the gene expression profiling using the microarray will catalyse the development of personalised treatment strategies for cancer patients based on the classification of molecular subtypes. (Fan & Ren, 2006) (Wang, Gerstein, & Snyder, 2009) As the amount of gene expression data generated from the microarray experiments is huge, the bioinformatics plays an essential role in acquiring, storing and analysing the data. There are a number of well-known bioinformatics tools already available to analyse the data generated in efficient ways and these tools are capable of dealing with messy biological data.

In this chapter, gene expression techniques used to perform downstream analyses will be discussed. Initially, the central dogma of molecular biology is discussed to introduce the concept of gene expression and provide general background. Next, microarray experiment, a popular technique to measure gene expression is discussed in detail. Microarray experiment generate numerical expression matrix and techniques for data quality checking and various pre-processing techniques are discussed. Thereafter, the statistical methods for microarray data analysis is discussed together

with the necessary computing environments. Next, RNA-Seq, a more recent technology for transcriptome profiling will be introduced and the pros and cons of microarray and RNA-Seq platforms will be discussed. Finally, the gene ontology enrichment analysis that associates identified differentially expressed genes to its enriched biological functions will be discussed.

2.2 Gene expression

The process in which the information of genes is used to synthesize its functional product, a protein is called gene expression. Gene expression processes are mainly handled by transcription and translation steps in a cell. The regulation of gene expression process refers to the control of the amount of functional product of genes produced and this is vital to cells, as the cells can control the production of gene products to respond to external stimulus, adapt to different environments, repair from the cell damage and undergo various cellular differentiations and morphogenesis. The regulation of genes expression controls the structures and functions of a cell. (Sánchez & de Villa, 2008) The next section covers the central dogma of molecular biology which describes the two-steps process, transcription and translation that forms the basis of gene expression process.

2.2.1 Central dogma of molecular biology

Protein is a product of genes that performs essential functions and forms various structures in cells. Proteins do not self-assemble, but they need sets of instructions to assemble and the instructions for protein sequence, structure and functions are contained within DNA. The DNAs essentially contain the complete genetic instructions that describes the functions and structures of a cell, and moreover, an organism. (Sánchez & de Villa, 2008) The central dogma of molecular biology is an explanation of the flow of genetic instructions from DNA to proteins introduced by Francis Crick in 1958. (Crick, 1970). The idea is that the genes contain genetic instructions to synthesize proteins and this is done in three steps namely, transcription, splicing and translation.

In eukaryotic cells, the DNA molecule is discovered in the form of a nucleoprotein complex named chromatin. Chromatin is a mass of genetic material composed of DNA and proteins that condense to form chromosomes during eukaryotic cell division. Chromatin is located in the nucleus of the cells. The primary function of chromatin is to compress the DNA into a compact unit that will be

less voluminous and can fit within the nucleus. Chromatin consists of complexes of small proteins known as histones and DNA. Histones help to organize DNA into structures called nucleosomes by providing a base on which the DNA can be wrapped around. A nucleosome consists of a DNA sequence of about 150 base pairs that is wrapped around a set of eight histones called an octamer. The nucleosome is further folded to produce a chromatin fibre. Chromatin fibres are coiled and condensed to form chromosomes. Chromatin makes it possible for a number of cell processes to occur including DNA replication, transcription, DNA repair, genetic recombination, and cell division. Chromatin can affect a gene's availability for transcription.

In transcription, a double stranded DNA unwinds, and each copy undergoes replication to produce identical copies of each helix. Each copies of helix are copied into a complementary mRNA helix. In splicing, the exons are joined together. Usually exons code for protein and the non-coding introns are removed by splicing. Proteins called transcription factors play a central role in regulating transcription. These proteins help determine which genes are active in each cell. For translation to happen, enzymes called RNA polymerase which makes a new RNA molecule from a DNA template must attach to the DNA of the gene. It attaches at the promoter. RNA polymerase can attach to the promoter only with the help of basal transcription factors. There are large class of transcription factors that control the expression of specific, individual genes. For instance, a transcription factor might activate only a set of genes needed in certain neurons. A transcription factor binds to DNA at a certain target sequence. Activators activate transcription by helping the general transcription factors or RNA polymerase bind to the promoter. Repressors represses transcription, for instance, they may get in the way of the basal transcription factors or RNA polymerase, making it so they can't bind to the promoter or begin transcription. The transcription start site is the location where transcription starts and it is located at 5'-end of a gene sequence and the core promoter includes the transcription start sites (TSS). When RNA polymerase binds to a promoter sequence at the TSS, the transcription starts and RNA polymerase would use one of the DNA strands to make a new complementary RNA molecule. (Sandelin et al 2004).

In translation, the amino acids are joined the form a sequence that results in protein. The amino acid sequences are determined by the order encoded in the mRNA known as a genetic code or a triplet codon. (Sánchez & de Villa, 2008) The figure 2.1 shows the whole process of gene expression.

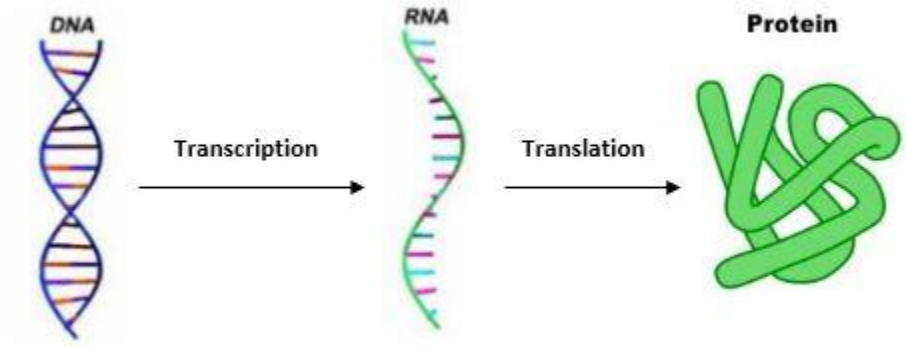


Figure 2.1 The schematic view of gene expression process based on the central dogma of molecular biology introduced by Francis Crick in 1958. The process starts with DNA replications and ends with translation that results in proteins (Crick, 1970).

Traditionally it was believed that a single gene results in a single protein which in fact, was not true due to various modifications such as alternative splicing and post-translational modifications in which one gene can result in several proteins. Following on from the understanding of central dogma of molecular biology, it has become important for biologists to understand what and how much mRNA is around so that they could identify the specific genes that display such expression intensity. For this reason, to quantitatively measure thousands of gene expression from experiments simultaneously, a high-throughput microarray technology has been developed. (Faiz & Burgess, 2012)

2.3 Microarray experiments and data analysis

The term microarray was defined by (Choudhuri, 2004) as "a high-throughput assay system which utilizes spatially ordered discrete, high-density arrangement of biologically important entities immobilized on a solid platform", where the entities usually are nucleic acids fragments, proteins carbohydrates etc. The study by (Pirrung, 2002) also defines microarray as "monolithic, flat surfaces that bear multiple probe sites, often hundreds and thousands, and each bear a reagent whose molecular recognition of a complementary molecule can lead to a signal that is detected by an imaging, most often fluorescence". The first microarray was an oligonucleotide-based microarray called GeneChip introduced by Affymetrix in 1996 and several types of microarrays have been developed and commercialized. (Faiz & Burgess, 2012) (Pirrung, 2002)

There are mainly two types of microarray technologies used in the biomedical research based on the material used as probe. First type is a complementary DNA known as cDNA and the second type is oligonucleotide microarray as explained in the above paragraph. The two microarrays usually have different probe and different target preparation (figure 2.2). In cDNA microarray, the probes are polymerase chain reaction (PCR) fragments that are amplified from the specific target sequences or clones from a cDNA library. cDNA microarray typically has been useful in the identification heterologous genes across species. The oligonucleotide microarrays are typically smaller (length of oligonucleotide microarray denoted by "25-mer") compared to the cDNA microarray (hundreds to thousands base pairs) and they do not utilize cDNA library, because their probes are synthesized on the chip and then immobilized on the platform. (Tarca, Romero, & Draghici, 2006) (Benjamini & Hochberg, 1995)

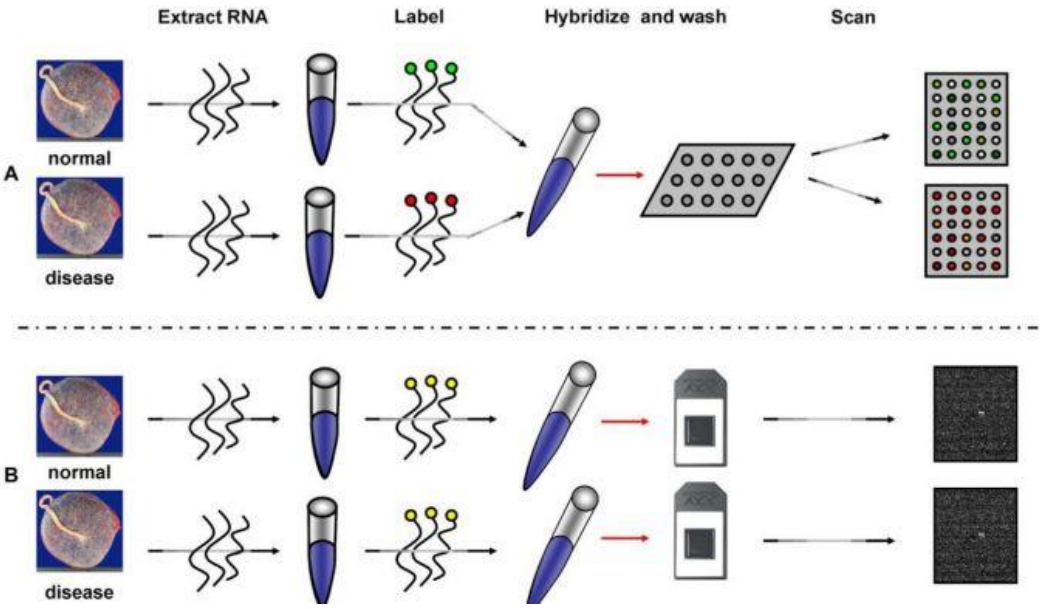


Figure 2.2 The schematic overview of probe array and target preparation for (a) cDNA microarrays: this is the two-channel microarray and (b) high-density oligonucleotide microarrays: this is a single channel microarray (Tarca et al., 2006).

Typically, the sample of DNA or RNA is separated from the biological sample goes through amplification. The samples are then labelled usually with the fluorescent dyes. Following the fluorescent labelling, the hybridization process happens between the complementary probes attached to the microarray surface and the isolated DNA or RNA samples. The microarray chips with hybridized double stranded molecules are rinsed to exclude non-specifically bounded

molecules and the fluorescent or radioactive images are recorded (known as image processing). The fluorescent images are produced by the label excitation by the laser scanner and they are captured by the camera. The fluorescent labelled hybridized image provides information about the locations of successful hybridization spot and the relative gene expression of a gene by displaying the intensity of the hybridized spot with specific probes. (Fan & Ren, 2006) (Pirrung, 2002) A higher intensity shows high gene expression level and the opposite is true as well. The microarray technologies thus, are capable of quantitatively measure thousands of gene expression levels simultaneously to examine the gene expression patterns after a specific type of experiments. Usually the gene expression of specific sequences is validated through RT-PCR. (Pirrung, 2002) Once, microarray experiments are performed, data must be analysed by incorporating several analysis procedures. Figure 2.3 shows typical microarray data analysis processes.

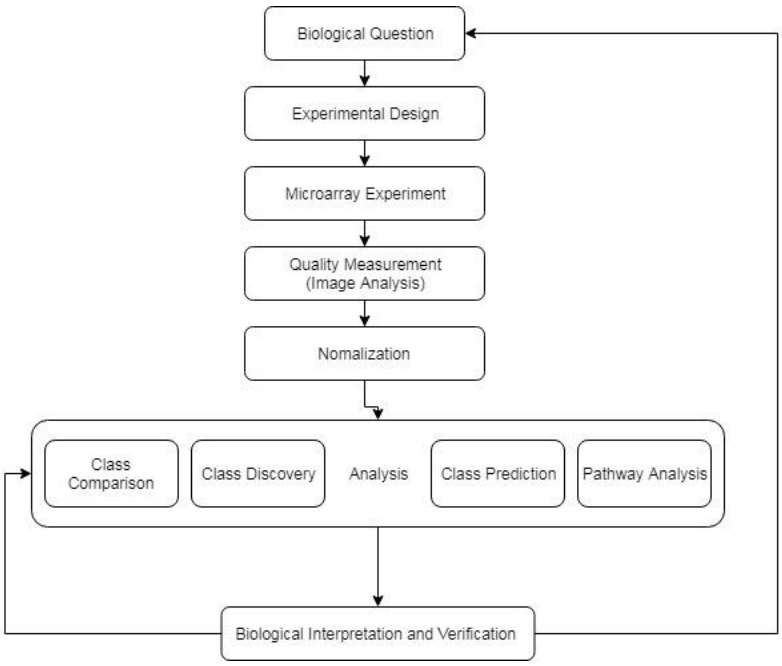


Figure 2.3 A schematic view of microarray data analysis processes.

The next few paragraphs cover several available computational tools to perform microarray analysis and gene expression data pre-processing steps including data quality checks and normalization techniques to transform the raw gene expression data to a suitable format that can be used for subsequent statistical analysis.

2.3.1 Methods and software

The growth of microarray applications in various research fields have increased the need for development of computational tools and analysis environment, new methods to model and analyse the data were required. The new tools were essential for processes such as obtaining and storing of the bulk of biological data generated. There are several commercial proprietary tools developed for bioinformatics analysis. These range from small software to much larger software for more complex data analysis such as Partek Genomics Suite that brings fast and optimized solutions for genomics data (Sánchez & de Villa, 2008). Commercial programs may have useful functions and tools optimized for biological the data, but they are usually very expensive, thus often an open source software may be the solution for bioinformaticians. There are obvious pros and cons of using an open source tools such as being prone to errors and not being maintained regularly, but open source software remains a good choice for biologists performing a gene expression analysis as there are several good open source software and supports available for this purpose. (Sánchez & de Villa, 2008) (Parmigiani, Garrett, Irizarry, & Zeger, 2003) The next few sections cover tools and analysis environment for gene expression analysis.

2.3.1.1 R programming language

R is a programming language designed for statistical computing and visualizations that is a free software supported by the R Foundation for Statistical Computing. The R programming language was created by Ross Ihaka and Robert Gentleman at the University of Auckland in New Zealand with the initial version release in 1995. It is available freely at https://www.r-project.org/. R has gained its popularity in the bioinformatics research areas due to a great support for statistical and graphical (for data visualizations) packages for complex biological data analysis, such as "ggplot2" and "dplyr" for visualizations and data manipulations respectively. (de Leeuw, 2009) R also supports object-oriented programming paradigm. R's one of the biggest advantages may be that it has a well-established system for package creation. There is a good amount of support by the community to create, test and distribute packages for different purposes and there are many packages created and distributed for biological analyses too such as the Bioconductor Project. (Gentleman et al., 2004) (de Leeuw, 2009)

2.3.1.2 Bioconductor project

The Bioconductor Project is an initiative for the collaborative development of computational biology and bioinformatics software. (Gentleman et al., 2004) The Bioconductor is an open source software project that provides a comprehensive support for genomic data analysis such as differential gene expression analysis and genome annotation etc. and it is based primarily on the R programming language. (Gentleman et al., 2004) One of the best approaches for genomics data analysis is incorporation of standard software with specific packages or libraries designed for microarray data analysis and the Bioconductor Project provides all the support for such purposes. The project (available at: https://www.bioconductor.org/) has grown extensively from 2001 and now, almost every microarray analysis technique has its own package. The R programming combined with the flexibility of the Bioconductor microarray analysis packages remain the chosen tool for statisticians and bioinformatician due to its capabilities to automate redundant analysis jobs in biological research and support for great visualizations and reporting tools. (Sánchez & de Villa, 2008) (Gentleman et al., 2004)

2.3.2 Pre-processing steps

Microarray data pre-processing is a crucial step that must be completed before proceeding with an actual statistical data analysis to remove any systematic variations and biases. (Allison, Cui, Page, & Sabripour, 2006) There are mainly two steps in microarray pre-processing. First, the data quality must be checked to determine if the data can be considered reliable. A typical high-throughput microarray data consist of huge numerical matrices and it is difficult to detect all the problems in the matrices by visual inspection, thus a specific data quality control procedure must be followed. Second, the quality-checked data must undergo normalization to correct for any technical artefacts. There are many normalization techniques developed and tested and selecting a specific normalization technique for microarray data analysis is crucial. (Sánchez & de Villa, 2008)

2.3.2.1 Data quality checks

The data quality check looks to examine and determine if the microarray data is reliable for subsequent statistical analysis. The quality assessment usually involves image analysis and several plots visualizations to detect any unreliable sources. Images of raw microarray data can be valuable

to detect irregular patterns such as scratches, rings, bubbles, shadows and lines. Sometimes it is difficult to check the images of all the raw microarray data, so methods such as signal-to-noise histograms or density plots of arrays that looks for significant multi-modality in the distribution can be used to detect poor quality arrays. RNA degradation plot can also be used to detect unreliable sources in the arrays. (Gillespie, Lei, Boys, Greenall, & Wilkinson, 2010) The figure 2.3 displays popular visualization techniques for detecting unreliable microarray data.

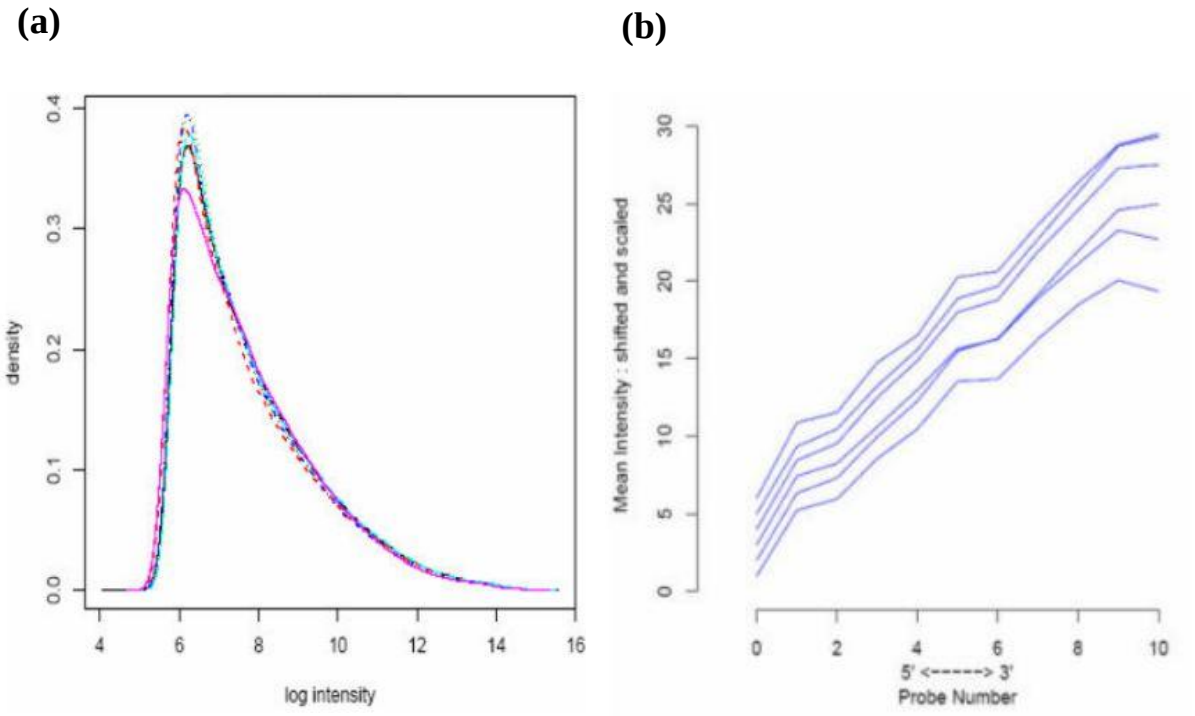


Figure 2.4 An example of diagnostic plots for the microarray samples. a) Density plot that checks for multi-modality. b) RNA degradation plot to detect the quality of RNA samples used to prepare microarray data (Sánchez & de Villa, 2008).

2.3.2.2 Normalization techniques

After checking the data quality, different technical artefacts must be removed, and the transformation applied to microarray gene expression data is referred to as normalization that aims

to adjust some individual hybridization intensities to balance them, from which meaningful biological analysis can continue. Typically, microarray analysis aims to explain the source of variation between RNA populations under different conditions (usually between diseased and normal) using their expression, to identify a gene or genes that may cause disease. The expression level observed, does not only include biological variation but it also includes variations introduced during sample preparation, array manufacture and the processing of arrays such as hybridization for example. These technical variation or biases, must be removed by data normalization to compare data from different arrays and different sources. (Irizarry et al., 2003) There are many normalization techniques developed for this purpose and some important normalizations are discussed in the next few sections.

2.3.2.2.1 Robust-multi array (RMA)

In high-density oligonucleotide microarrays provided by Affymetrix GeneChip, the most frequently used method is Robust-Multichip Average (RMA) normalization. The oligonucleotide microarray probes consist of 25 base pairs in length with two types of probes namely, perfect match (PM) and mismatch (MM). The PM probes are reference probes that match the target sequence perfectly, but MM probes differ from the reference sequence by one single base and the intensities from each probe in a probe-set are combined to provide an expression measure. (Irizarry et al., 2003) (Bolstad, Irizarry, Åstrand, & Speed, 2003) In RMA normalization, a robust average of $\log_2$ background corrected PM intensities are used to make an estimate. The RMA normalization typically consists of 3 steps. First, a probe level background adjustment is made. Second, quantile normalization is made and lastly, summarization of all the values of probe related to one gene is made. There are many other normalization methods such as MAS5.0 developed by the Affymetrix manufacturer, but RMA normalization is preferred as they tend to generate less false positives. (Sánchez & de Villa, 2008)

2.3.2.2.2 Quantile normalization

In some microarray experiments, numerical gene matrices from different arrays are combined. For example, an experiment could contain Affymetrix arrays, HG-U133A and HT_HG-U133A and before the gene matrices from two different arrays can be combined for a bigger analysis, a method must exist to remove a noise derived from different physical arrays. Quantile normalization

(typically known as within slides normalization) is a method that achieves consistency between different arrays by making the distribution of probe intensities from different arrays uniform. (Sánchez & de Villa, 2008) (Irizarry et al., 2003) The boxplots are usually a very useful tool to visualize the effects of different normalization methods on the microarray data.

2.3.2.3 Batch effects

Once the data quality has been checked and technical artefacts are removed through normalization, it is important to check if the transcriptomic dataset contains any batch effect. Batch effects are the systematic errors introduced when samples are handled in multiple batches in a laboratory. Through careful design of experiments, batch effects may be reduced, but it cannot be eliminated unless the whole experiment was done in a single batch. (Chen et al., 2011) Batch effects in gene expression from the vast expression dataset may typically come from cell culture conditions and other uncontrollable variables such as chip type, platform, laboratory, technicians, storage and shipment conditions and protocols etc. and these factors confound gene expression variation due to an actual experimental condition for instance, the test of different drugs. This technical bias must be detected and removed otherwise this could have a serious effect on downstream microarray analysis results. (Reese et al., 2013) There are number of methods available to adjust for batch effects.

A common method to detect batch effects in the transcriptomic data is the use of principle component analysis (PCA), an unsupervised learning algorithm used for dimensionality reduction and interpretation. PCA aims to identify the 'combination of conditions that explain the greatest variation in the data' (Yang et al., 2008). Typically, in PCA, the first two principle components with each sample coloured according to their respective batch number are plotted. If there is a clear separation of colours in the PCA plot, it probably indicates that the greatest variation comes from the batch effect and this must be removed. Once the batch effects are removed and the first principle components are re-plotted, the color-coded samples will cluster together, since now the greatest source of variation in gene expression may be anything else than the batch effects. (Reese et al., 2013) There are several ways to remove the batch effects, but one of the most popular methods is to utilize the 'ComBat' function available from the SVA package through the Bioconductor Project (section 2.3.1.2). The 'ComBat' function requires a batch vector as one of

its input and it will return a batch-corrected gene expression matrix. After the global normalization of the expression data, detection and the removal of batch effects is a crucial step in building a gene expression matrix that can be used for downstream statistical analysis and this step should never be underestimated.

2.3.2.4 Gene annotation

The numerical gene matrix obtained from the pre-processing, typically has the Affymetrix probe set identifiers in its rows and the sample's experimental condition in its columns. The probe set identifiers must be mapped with their corresponding gene symbols for the identification of differentially expressed genes during statistical analysis and visualizations. A well-known bioinformatics packages from the Bioconductor Project (section 2.3.1.2) are usually utilized for gene symbol annotations. The packages such as, "hgu133a.db" and "hthgu133a.db" can be used to annotate the Affymetrix probe set identifiers of chips like HG-U133A and HT_HG-U133A. Once the probe identifiers are mapped with their corresponding gene symbols or ENTREZ gene identifiers for subsequent statistical analysis to identify genes of interest or to perform gene ontology analysis. (Gentleman et al., 2004)

2.3.3 Differential expression statistical analysis

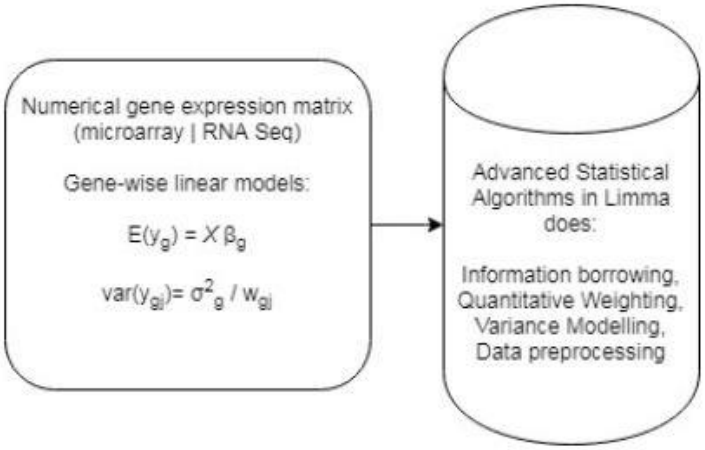
After microarray data has been pre-processed following several standard procedures, an important statistical question is to detect and select genes that are differentially expressed between the experimental and control samples. When designing a microarray data analysis experiment, it is crucial to consider the presence of biological replicates, sometimes known as within-array replications. Biological replicates are samples of same genetic material that have been tested with a specific experimental condition, that are expected to show similar gene expression pattern. Biological replicates are used in microarray data analysis to improve the precision of statistical methods designed to identify differentially expressed genes. (Fan & Ren, 2006)

Differential expression analysis is a statistical inference method to identify genes whose expression levels are significantly different between two groups of experiments. (Selvaraj & Natarajan, 2011) The analysis starts with selection of proper statistical test. Some of the most

widely used methods include, Significance Analysis of Microarrays (SAM), one-sample and two-sample t test, F-statistics and an empirical Bayes method. (Fan & Ren, 2006) In the next few sections, one of the most popular methods for microarray differential expression analysis is described.

2.3.3.1 Linear modelling for microarrays

LIMMA is a popular Bioconductor software package that provides comprehensive set of tools for microarray gene expression analysis and the package has incorporated several statistical principles for large scale expression analysis. Initially, several matrices of gene expression values are created known as design matrix and contrast matrix that represent RNA targets and experimental conditions respectively. The rows of matrices usually represent a gene and the columns represent the experimental conditions. Once, experimental matrices are constructed, a linear model is fitted to each row of gene expression data to compute test statistics including t-statistics and F-statistics. (Selvaraj & Natarajan, 2011) In this way, a complex experimental designs and hypotheses can be tested and the strength between the gene-wise models are borrowed to increase the reliability of statistical test even with the small number of samples. (Smyth, Ritchie, Thorne, & Wettenhall, 2005) Figure 2.5 highlights the statistical principles utilized by LIMMA analysis.



Powerful inference for differential expression analysis can be delivered. Parameters used for gene prioritization and gene set testing include: $\boldsymbol{\beta_g}$ **and** $\boldsymbol{s^2_g}$

Figure 2.5 Schematic of major statistical principles of LIMMA analysis. Initially, for each gene g, its gene expression values and a design matrix X relate to coefficients of interest (βg). Typically, empirical Bayes methods are used to obtain posterior variance estimator (s^{2*}_g) to facilitate gene-wise information borrowing process. LIMMA also utilizes observation weight to allow for data quality variations, variance modelling to count for technical or biological differences that are present and pre-processing methods to remove biases and noises present in data. The combination of these statistical principles all contribute to the improvement of statistical inference for genes and gene sets in microarray data analysis with small number of samples (Smyth et al., 2005).

After linear model fitting, the standard errors are moderated using a simple parametric empirical Bayes model that borrows strength between genes to moderate the residual variances. This method after linear fitting computes a moderated t-statistics and log-odds of differential expression for each contrast of each gene. Empirical Bayes method that incorporate posterior variance estimators effectively increased the degrees of freedom in which the gene-wise variances are estimated, and it has been proved that this method was particularly useful in microarray data analysis with small sample size in terms of reliable statistical inference. (Smyth et al., 2005) (Reiner, Yekutieli, & Benjamini, 2003) Empirical Bayes also computes a moderated F-statistics that combines the t-statistics for all the contrasts for each gene into an overall test of significance for the gene. Typically, the moderated F-statistics tests for non-zero in contrasts for the gene and the p-values are generated. In an experiment with a numerous contrast, it is usually recommended to select differentially expressed genes based on moderated F-statistics and significance of the contrasts for the genes. This is especially advantageous as the number of tests can be cut down to reduce the amount of adjustment for multiple testing. (Smyth et al., 2005) In overall, LIMMA package available through Bioconductor project provide a comprehensive set of powerful tools that can be used to detect differentially expressed genes in reliable manner with improved statistical power and accuracy.

2.3.3.2 False discovery rate control

In a microarray statistical analysis, controlling the false discovery rate is one of the most important steps before the selection of the differentially expressed genes. A microarray data analysis involves simultaneous hypothesis testing of several thousands to tens of thousands of genes and the probability of finding a false positive increase when the number of genes tested increases. To correct for this multiple hypothesis testing problems and generate less false positives, a false

discovery rate usually known as the rate of 'type I errors' must be corrected. (Reiner et al., 2003) The most popular form of adjustment is introduced by Benjamini and Hochberg that controls the false discovery rate. R's base statistical packages provide p-value adjustment functions to control the false positives. There are several options such as Bonferroni adjustment and FDR (known as Benjamini Hochberg adjustment in R) adjustment, and FDR method tends to be most powerful as the false discovery rate is a less stringent condition than family-wise error rate. (Benjamini & Hochberg, 1995) Usually the p-values obtained from the moderated t-statistics after adjustment for multiple testing are used to select differentially expressed genes along with other criteria such as the log fold change values. Following the p-value adjustment, a p-value threshold is selected for identification of differentially expressed genes. If for example, p-value threshold of 0.05 has been selected, all the genes with p-values below 0.05 are selected, meaning that the expected proportion of false discoveries in the selected group should be less than 5%. (Reiner et al., 2003) Researchers can set p-values based on how strictly they want to control the false positives but in a typical biomedical researches p-value threshold is usually set at 0.05. (Dalman, Deeter, Nimishakavi, & Duan, 2012) By controlling the false discovery rate, a more reliable set of genes from differential expression analysis can be selected for subsequent clustering or functional analysis.

2.3.3.3 Clustering and visualization techniques

Clustering of gene expression data usually falls into class discovery problem. In class discovery, genes that are similarly expressed are grouped together using several clustering methods and the results are correlated to biology for interpretations. If the genes are co-regulated and expressed (up or down regulation) simultaneously, they may be functional related genes and the functionally related genes can be clustered into groups. Clustering, like principal component analysis can reduce the dimensionality of microarray data. (Wang et al., 2009) Clustering can be applied to both genes and experimental samples and when the experimental samples (usually columns from the matrix) are clustered for instance, the drugs with similar effects can be identified. Hierarchical clustering technique has been most popular clustering method to identify a partition of the experimental samples more than the genes as the size of samples are less than the size of genes and it is usually difficult to interpret the dendrogram produced by the gene clustering. (Wang et

al., 2009) In hierarchical clustering, each cluster are divided into smaller clusters that look like a tree structure called dendrogram. In agglomerative hierarchical clustering, the single-gene clusters are joined by the closest clusters until all the genes have been joined into the original cluster of genes. There are several inter-cluster distance approaches used by hierarchical clustering including, single linkage, complete linkage, average linkage and centroid linkage. (D'haeseleer, 2005) Once the genes have been hierarchically clustered using some linkage methods, their expression patterns are displayed using visualization tools. Heatmaps are very useful visualization tool to monitor gene expression pattern overall. Heatmaps consist of rectangular array of coloured blocks that represent the expression level of genes. Typically, red maps represent up-regulation and green maps represent down-regulation expression, but many other colour choices can be used too. Through heatmaps, the gene expression intensities can be visualized to examine how strongly has the gene been expressed under different experimental conditions and identify similarly expressed group of genes. (Wang et al., 2009)

2.3.3.3.1 Weighted gene co-expression network analysis

In bioinformatics applications, co-expression analysis is a powerful tool to identify genes involved in the same biological process. Genes that display similar expression patterns are likely to be involved in same biological functions or pathways and these genes are known as co-expressed genes. Co-expressed genes are useful in identifying candidate genes for metabolic pathway enzymes or transcription factors that regulate them. (Oldham et al., 2008) Bioconductor offers a comprehensive software package for performing various aspects of weighted gene correlation network analysis known as Weighted Gene Co-expression Network Analysis (WGCNA). (Langfelder & Horvath, 2008) WGCNA package provides several functions to perform weighted correlation network analysis, module detection, gene selection, visualization and interfacing with external software such as ‘Cytoscape’ (Shannon et al., 2003). Typically, WGCNA aims to identify modules (clusters) of highly correlated genes that have high topological overlap which is a pair-wise measure that displays two genes’ co-expression relationships with other genes. (Oldham et al., 2008) High topological overlap displayed by the genes show that they are highly correlated in the network. WGCNA uses hierarchical average linkage clustering method to cluster highly correlated genes into a co-expression module. (Langfelder & Horvath, 2008) The figure 2.6 presents an overview of WGCNA procedures.

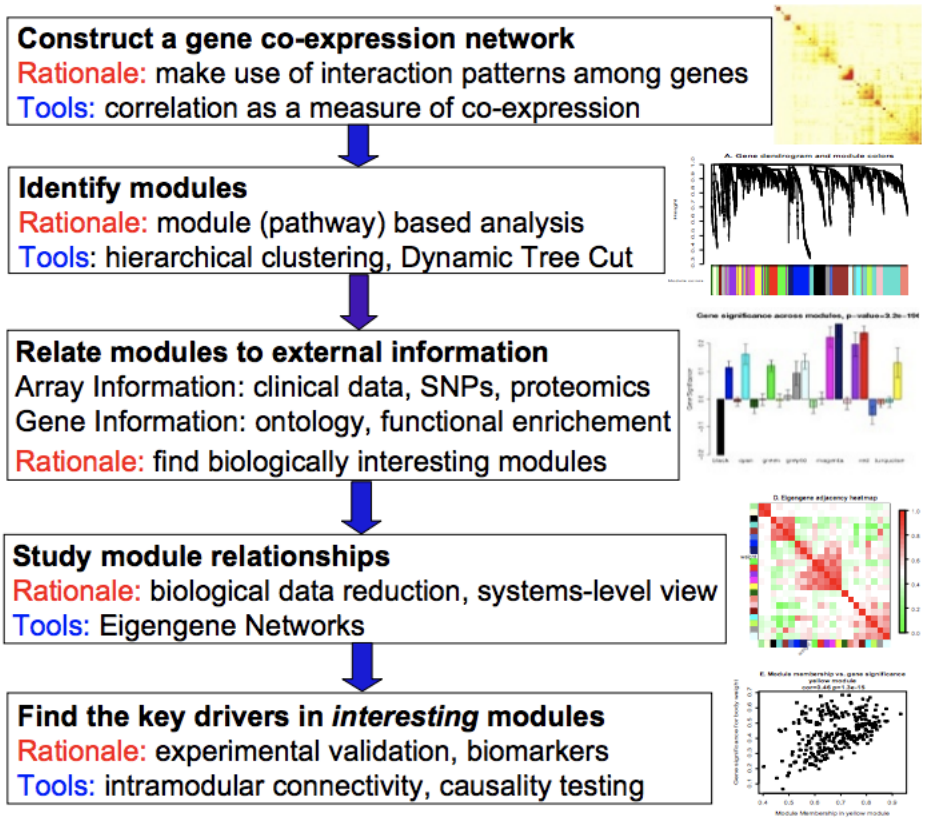


Figure 2.6 Schematic overview of WGCNA procedures. A gene co-expression networks are constructed to identify modules of co-expressed genes using hierarchical average linkage clustering method and dynamic tree cut. Significant modules can be selected, and they can be related to their biological traits information. A subsequent functional enrichment analysis can be performed using the gene modules (Langfelder & Horvath, 2008).

Once different co-expression modules are identified using WGCNA, a gene ontology enrichment analysis or biological pathway analysis can be performed to identify biological functions or pathways that are significantly enriched by the co-expression modules as seen in the last step of figure 2.6. WGCNA approach has been widely used in different cancer and brain researches (Oldham et al., 2008) for candidate biomarker and therapeutic target discoveries.

2.4 Biological interpretations

The usual goal of microarray experiments is to identify a set of differentially expressed genes between two or more conditions. (Sánchez & de Villa, 2008) Typically, the analysis will produce a long list of genes that had passed several criteria such as p-value threshold and the log fold-change. These genes are known to be statistically significant and the researchers can associate these genes to a specific biological functions or pathways that the genes are involved in. There are databases designed for this purpose namely, Gene Ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG).

The gene enrichment analysis is a high-throughput approach that is used to associate the list of identified genes to their biological functions or processes and the pathways using databases such as GO and KEGG. (Huang, Sherman, & Lempicki, 2008) The gene set enrichment analysis is usually performed to identify if the differentially expressed genes have been significantly enriched or impoverished in a given category of biological processes. During the analysis, a numerical variable such as p-values are used to rank the enrichment score based on the presence or absence of the genes in the specific biological processes. A Kolmogorov-Smirnov test is usually used to make decisions about the list of genes' representations in the biological process. (Sánchez & de Villa, 2008)

There are several high-throughput enrichment analysis tools developed such as Onto-Express, GoMiner, EASE and DAVID. The GO enrichment analysis can also be done through the WGCNA package which is a co-expression analysis package available from the Bioconductor Project (section 2.3.2.1). (Huang et al., 2008) The tools available for enrichment analysis may have different functionalities and features but the underlying concept of the tools is the same. The figure 2.9 describes the 3 layers of gene enrichment analysis workflow.

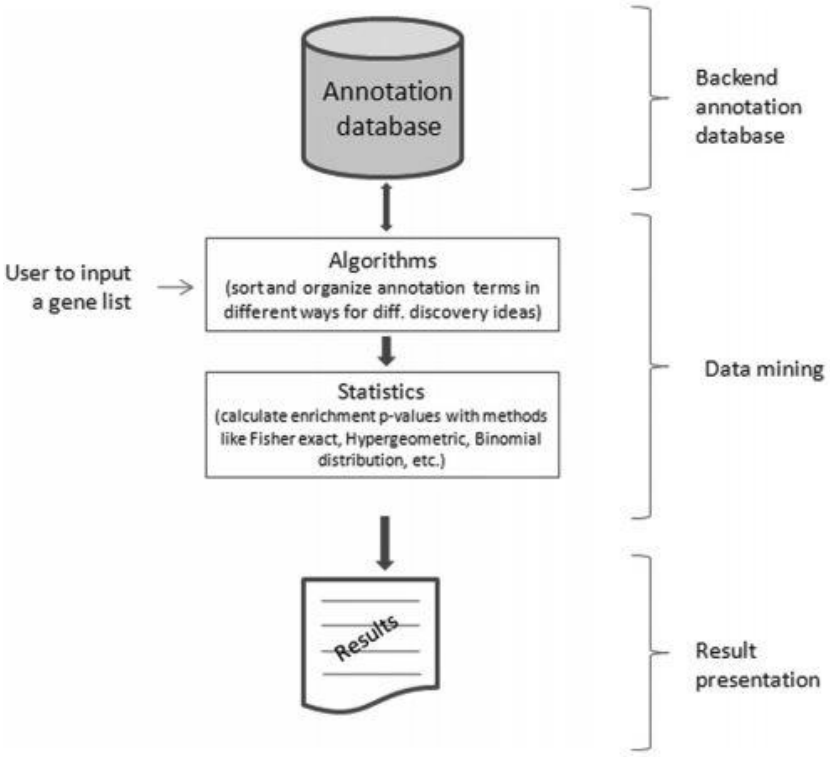


Figure 2.7 The schematic overview of the infrastructure of gene enrichment analysis tools. There are three major layers to this workflow. In the first layer (backend annotation database), annotation databases such as GO and KEGG are available. In the second layer (data mining), the user inputs the list of genes and goes through several statistical tests to generate results such as enrichment score using p-values. Each layer has a great influence in the result (Huang et al., 2008).

Due to the complex nature of biological data-mining, the enrichment analysis of a large list of differentially expressed genes is still regarded as an exploratory data analysis than a pure statistical analysis. (Huang et al., 2008) Typically, the best biological conclusions can be drawn, when the researcher's thorough knowledge of biology, integration of useful biological databases and incorporation of computing algorithms and statistical methodologies are combined. The enrichment analysis tools must still be improved, and it is crucial that the researchers must take into consideration of the biological facts and statistical analysis results together to interpret the results from functional analysis. (Huang et al., 2008)

2.5 Challenges with microarrays

Microarray technology has been very useful in drug discovery and disease researches, but data mining with microarray still presents several challenges. The sample size in microarray analysis is usually small and this may be problematic. The number of genes may range from 1,000s to 10,000s while the size of samples ranges from 10s to 100s only. The number of samples in microarray analysis due to difficulty in the collection of microarray samples, could significantly increase the finding of false positives that are due to chances and not biological variation. (Piatetsky-Shapiro & Tamayo, 2003) The amount and quality of RNA isolated from microarray experiments remains a major challenge, due to the complex nature of tissue samples obtained. This limitation has created false microarray data that were generated from degraded mRNA sources and the experiments had to be replicated several times to eliminate errors. (Russo, Zegar, & Giordano, 2003) Currently, microarray technology is limited by the tissue samples, arrays and analysis methodologies. Due to these reasons, microarray data analysis pipeline requires more robust and validated models to accurately analyse noisy microarray data.

2.6 RNA-Seq

RNA-Seq is a recent gene expression profiling method that uses deep-sequencing technologies. There are several advantages of using RNA-Seq profiling method over existing methods, although, it is still under an active development. (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008) RNA-Seq technology can be useful in expression profiling of non-model organisms with genome sequences that have not yet been determined, whereas existing hybridization-based arrays are limited to existing genome sequences. It is also possible to precisely identify location of transcription boundaries to a single-base resolution. (Sánchez & de Villa, 2008) RNA-Seq typically also has very low background signal with a large dynamic range of expression levels for transcript detection. RNA-Seq also provide an accurate quantification of expression levels. (Sánchez & de Villa, 2008) As with any other array technologies, RNA-Seq also has limitations of efficiently storing, retrieving and processing of large quantities of expression data, but in overall RNA-Seq offers far more accurate and precise quantification of transcriptional expression measurements compared to classical hybridized-based arrays like microarrays. The expression

profiling technology must be carefully chosen by the researcher accordingly, to match the purpose of the research as each technology has its own advantages and disadvantages.

2.7 References

1. Allison, D. B., Cui, X., Page, G. P., & Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. Nature reviews genetics, 7(1), 55.
2. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological), 289-300.
3. Bolstad, B. M., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19(2), 185-193.
4. Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., & Liu, C. (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PloS one, 6(2), e17238.
5. Choudhuri, S. (2004). Microarrays in biology and medicine. Journal of biochemical and molecular toxicology, 18(4), 171-179.
6. Crick, F. (1970). Central dogma of molecular biology. Nature, 227(5258), 561.
7. D'haeseleer, P. (2005). How does gene expression clustering work? Nature biotechnology, 23(12), 1499.
8. Dalman, M. R., Deeter, A., Nimishakavi, G., & Duan, Z.-H. (2012). Fold change and p-value cutoffs significantly alter microarray interpretations. Paper presented at the BMC bioinformatics.
9. de Leeuw, J. (2009). R Programming for Bioinformatics. Journal of Statistical Software, 29(1), 1-2.
10. Faiz, A., & Burgess, J. K. (2012). How can microarrays unlock asthma? Journal of allergy, 2012.
11. Fan, J., & Ren, Y. (2006). Statistical analysis of DNA microarray data in cancer research. Clinical Cancer Research, 12(15), 4469-4473.
12. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., . . . Gentry, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome biology, 5(10), R80.
13. Gillespie, C. S., Lei, G., Boys, R. J., Greenall, A., & Wilkinson, D. J. (2010). Analysing time course microarray data using Bioconductor: a case study using yeast2 Affymetrix arrays. BMC research notes, 3(1), 81.

14. Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research, 37(1), 1-13.

15. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 4(2), 249-264.

16. Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics, 9(1), 559.

17. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods, 5(7), 621.

18. Oldham, M. C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., & Geschwind, D. H. (2008). Functional organization of the transcriptome in human brain. Nature neuroscience, 11(11), 1271.

19. Parmigiani, G., Garrett, E. S., Irizarry, R. A., & Zeger, S. L. (2003). The analysis of gene expression data: an overview of methods and software The analysis of gene expression data (pp. 1-45): Springer.

20. Piatetsky-Shapiro, G., & Tamayo, P. (2003). Microarray data mining: facing the challenges. ACM SIGKDD Explorations Newsletter, 5(2), 1-5.

21. Pirrung, M. C. (2002). How to make a DNA chip. Angewandte Chemie International Edition, 41(8), 1276-1289.

22. Reese, S. E., Archer, K. J., Therneau, T. M., Atkinson, E. J., Vachon, C. M., De Andrade, M., . . . Eckel-Passow, J. E. (2013). A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. Bioinformatics, 29(22), 2877-2883.

23. Reiner, A., Yekutieli, D., & Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics, 19(3), 368-375.

24. Russo, G., Zegar, C., & Giordano, A. (2003). Advantages and limitations of microarray technology in human cancer. Oncogene, 22(42), 6497.

25. Sánchez, A., & de Villa, M. (2008). A tutorial review of microarray data analysis. Universitat de Barcelona.

26. Sandelin, A., Alekma, W., Engstrom, P., Wasserman, W. W., & Lenhard, B (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids research, 32(suppll_1), D91-D94.

27. Selvaraj, S., & Natarajan, J. (2011). Microarray data analysis and mining tools. Bioinformation, 6(3), 95.

28. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11), 2498-2504.

29. Smyth, G. K., Ritchie, M., Thorne, N., & Wettenhall, J. (2005). LIMMA: linear models for microarray data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Statistics for Biology and Health.

30. Suárez, E., Burguete, A., & McLachlan, G. J. (2009). Microarray data analysis for differential expression: a tutorial. Puerto Rico health sciences journal, 28(2).

31. Tarca, A. L., Romero, R., & Draghici, S. (2006). Analysis of microarray experiments of gene expression profiling. American Journal of Obstetrics & Gynecology, 195(2), 373-388.

32. Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews genetics, 10(1), 57.

33. Yang, H., Harrington, C. A., Vartanian, K., Coldren, C. D., Hall, R., & Churchill, G. A. (2008). Randomization in laboratory procedure is key to obtaining reproducible microarray results. PloS one, 3(11), e3724.

3 GLYCOBIOLOGY AND GLYCOSYLTRANSFERASES

3.1 Introduction to glycobiology

Glycobiology by its definition is the study of the structure, chemistry, biosynthesis and biological functions of glycans known as carbohydrates and their derivatives. Glycobiology is a small but fast-growing field in biology with its relevance in biomedicine, biotechnology and cancer researches. Proteomics, which is the study of protein expression, modifications, structures and functions has expanded our knowledge of proteins but typically in eukaryotic cells, proteins undergo posttranslational modifications. One of the major posttranslational modifications is glycosylation whereby glycans essential for cell's viability are attached to the protein structures. Oligosaccharides can be covalently attached to proteins in two ways namely, N-glycosidic linkage between N-Acetylglucosamine (GlcNAc) and an asparagine residue or an O-glycosidic linkage between N-Acetylglucosamine (GalNAc) and the hydroxyl group of serine or threonine. (Stanley, Schachter, & Taniguchi, 2009) There are many functions of glycans in cell but one of its main functions is to be involved in intercellular interactions where cell to cell and cell to pathogen contact occurs through glycan-protein interactions. Glycan to protein interactions occur through glycosylation which is modulated by glycosyltransferase enzymes. There are currently more than 150 different glycosyltransferase enzymes identified but there are at least double this number to be characterized. In this short chapter, an overview of carbohydrates, glycosylation and glycosyltransferase will be presented and their roles in breast cancer will be discussed to give an idea of how the applications of glycobiology can be useful in cancer researches and the development of potential cancer biomarkers.

3.1.1 Carbohydrates

There are two major classes of molecules required to build a cell, namely, lipids and carbohydrates. Carbohydrates are biomolecules that are involved in numerous biological processes including, oligomerization, protein folding, cell to cell recognition and response and host-pathogen interactions. (Berninsone, 2005) Carbohydrates function as an intermediate in energy generation and signalling effectors, structural components and recognition markers and major posttranslational modifications of proteins are also encompassed by carbohydrates (Stanley et al.,

2009) With this, it is apparent how a small number of carbohydrate related genes can create complexities in development, growth and functioning of a cell in a biological system.

3.1.1.1 Monosaccharides

Monosaccharides, which are the basic units of carbohydrates are the basic structural units of glycans which is a compound consisting of large number of monosaccharides that are linked via glycosidic linkages. Monosaccharides cannot be hydrolysed into a simpler form and there are two types of monosaccharides according to the position of carbonyl group in the carbon chain. If the carbonyl is attached at the end of carbon chain, they are called aldose of if it is attached to the inner carbon, it is called ketose (Stanley et al., 2009). In its natural state, a monosaccharide exists in either open-chain or ring formations (figure 3.1).

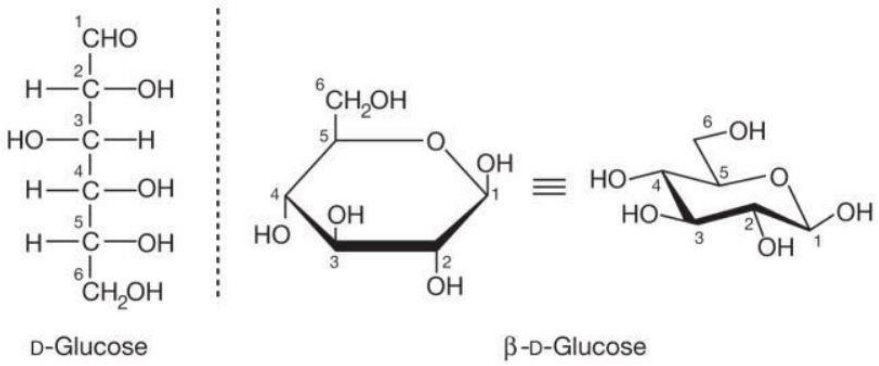


Figure 3.1 Open chain and ring formation of glucose. Typically, changes in the orientation of hydroxyl groups around specific carbon atoms create a new molecule such as galactose that is the C-4 epimer of glucose (Stanley et al., 2009).

Monosaccharides can be attached to another residue via a glycosidic linkage that involves the hydroxyl group of anomeric centres, creating α linkages or β linkages. The two different linkages confer different biological functions of glycans. (Berninsone, 2005) A glycoconjugates are a compound that consists of more than one monosaccharides or oligosaccharide units that are covalently bonded to a non-carbohydrate moiety and usually, the glycans constitute a major portion of glycoconjugates. There are several hundred distinct monosaccharides occurring

naturally but only a few of these are found in animal glycans. The table 3.1 shows the list of common monosaccharide units of glycoconjugates.

Table 3.1 List of most common monosaccharides units of glycoconjugates.

Name	No. of Carbons	Description
Pentose	5	Neutral sugars, e.g. D-xylose
Hexose	6	Neutral sugars, e.g. D-glucose (Glc), D-galactose (Gal), D-mannose (Man)
Hexosamines	6	Hexoses with amino groups at the 2-position, e.g. N-acetyl-D-glucosamine (GlcNac) and N-acetyl-D-galactosamine (GalNac)
Deoxyhexoses	6	Neutral sugar without the hydroxyl group at the 6-position
Uronic Acids	6	Hexose with a negatively charged carboxylate at the 6-position, e.g. D-glucuronic acid (GlcA)
Sialic Acids	9	N-acetylneuraminic acid (Neu5Ac)

This limited set of monosaccharides can undergo several different modifications, and this increases their diversity and mediates their biological functions. The hydroxyl groups of monosaccharides thus undergo modifications such as, phosphorylation, methylation, O-acetylation, or fatty acylation. (Stanley et al., 2009)

3.1.1.2 Classes of glycoconjugates and glycans

The classes of glycans naturally occurring are defined by the type of linkage to the protein or lipids. Proteins can also have glycoconjugates covalently attached via N or O linkages. (Pinho & Reis, 2015) N-glycans are oligosaccharides chain that has asparagine residue covalently attached to a polypeptide chain. An O-glycans have a linkage between sugar chain and polypeptide via N-acetylgalactosamine (GalNac) to a hydroxyl group of a serine or threonine residue. Mucin is a

glycoprotein that has many O-glycans that are grouped together. (Stanley et al., 2009) Figure 3.2 displays different classes of glycoconjugates occurring commonly in mammalian cells.

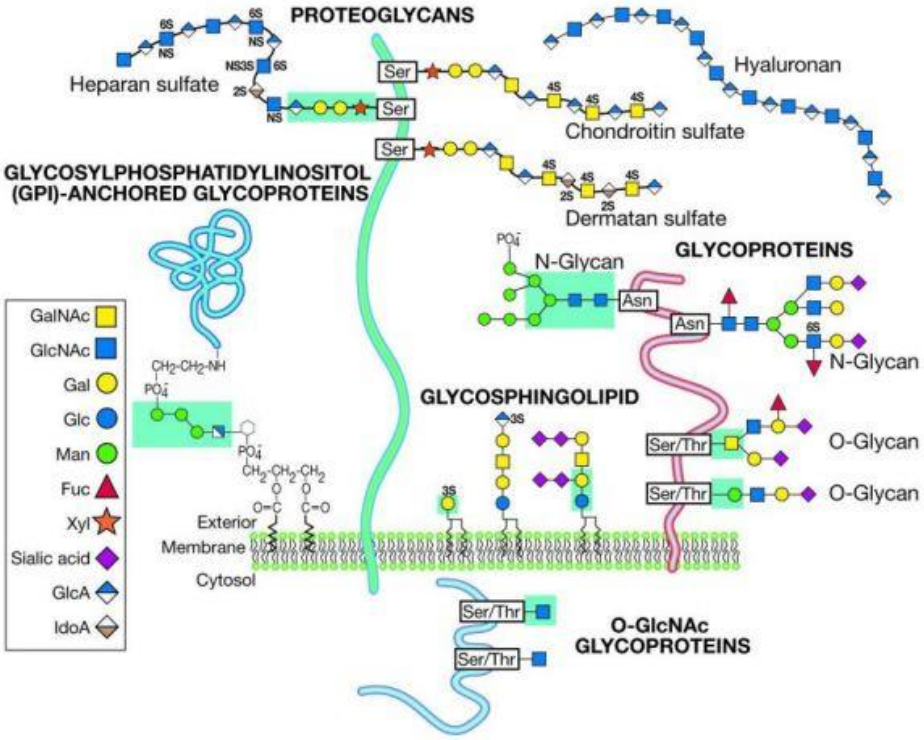


Figure 3.2 Different classes of common glycoconjugates occurring in mammalian cells. Glycosphingolipids typically appear on the outer leaflet of the cell plasma membrane. These glycans can be modified by terminal sialic acids. Saccharides can be covalently attached to a polypeptide backbone via N-linkage to Asn or O-linkage to Ser/Thr to produce a glycoprotein (Iha & Yamada, 2013).

There are other major glycoconjugates including the proteoglycans and glycosphingolipids (figure 3.2). Proteoglycans usually have glycosaminoglycan (GAG) such as keratan sulphate in its component. Glycosphingolipids are composed of glycans that are linked to a lipid ceramide and these molecules have a series of neutral structures and gangliosides that carries sialic acids which is known to regulate the receptor tyrosine kinase (RTK) signalling. (Pinho & Reis, 2015) (Iha & Yamada, 2013) Glycan chain structures unlike protein sequences are not determined and characterized directly by the genome and these are the secondary gene product. A few number of genes in the human genome encode for enzymes and transporters directly involved in the

biosynthesis and assembly of glycans through posttranslational modifications of proteins and glycosylation by enzymatic processes. (Stanley et al., 2009) This results in glycans having many different combinatorial ways they can assemble and for this reason, it is not possible to understand all the structures and pathways to predict the structures of glycans in a given cell. (Stanley et al., 2009) Additionally, environmental changes can result in significant changes of glycan structures and it is the complex and dynamic nature of glycosylation that makes the research field of glycobiology more difficult to conduct than any other topics such as nucleic acids or proteins.

3.1.2 Glycosylation

Glycosylation is a biological reaction or enzymatic process that attaches carbohydrates to proteins, lipids and other organic molecules. (Kailemia, Park, & Lebrilla, 2017) Glycosylation is the most common form of posttranslational modification and in this reaction, generally, saccharide units are covalently attached to the target proteins and sequential elongation of the branches proceed. (Hakomori, 2002) Glycosylation is typically proceeded by numerous glycan modifying enzymes including, glycosyltransferases and glycosidases resulting in various complex carbohydrates like, glycoproteins, glycolipids and proteoglycans. The rate of glycosylation for a given protein is determined by the presence and the frequency of glycosylation sites of proteins as well as the transcriptional levels and activities of glycosylation enzymes within the cell. (Pinho & Reis, 2015) (Munkley & Elliott, 2016)

The glycosylation in a cell produce two types of glycans namely, N-glycans and O-glycans (section 3.1.1.2) In N-glycosylation, an oligosaccharide precursor is transferred in groups to nascent proteins in the endoplasmic reticulum (ER). (Li, Song, & Qin, 2010) A subsequent reaction happen in the ER after the transfer of precursor that includes glucose removal and addition in a cycle that contributes to the protein folding. Additionally, N-glycan chains can be even more diversified in the Golgi apparatus. (Reis, Osorio, Silva, Gomes, & David, 2010) O-glycosylation is the process where glycans get O-linked to a serine or a threonine residue (figure 3.2), and the rate of this type of glycosylation in glycoproteins is high on a membrane-bound mucin. In O-linked glycosylation, N-acetylglucosamine (GalNAc), is transferred to serine and threonine residues from a sugar donor UDP-GalNAc, and this is usually modulated by UDP-GalNAc-polypeptide N-acetylgalactosaminyl-transferases (ppGalNAc-Ts). Altered transcriptional levels of ppGalNAc-Ts

may result in mucin O-glycosylation changes that is present during the malignant transformation in a cell. (Reis et al., 2010) In the next section, the enzyme involved in typical glycosylation processes will be discussed such as glycosyltransferase enzymes in terms of its types and roles in general.

3.1.3 Glycosyltransferases

The glycan biosynthesis is primarily regulated by the glycosyltransferase enzymes that catalyses the formation of glycosidic linkage to create carbohydrates, glycoside, oligosaccharides or polysaccharides. These enzymes assemble monosaccharides moieties of simple nucleotide sugar donor substrate (UDP-Gal, GDP-Fuc or CMP-Sia) into linear and branched glycan chains by transferring glycosyl residues from nucleotide-activated phosphoglycosides to other carbohydrates. (Paulson & Colley, 1989) Figure 3.2 shows a typical glycosylation reaction.

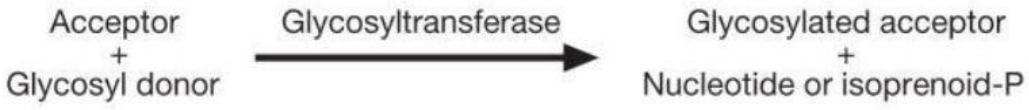


Figure 3.3 Glycosylation reaction showing activity of glycosyltransferase enzyme. A glycosyl donors include nucleotide sugar and dolichol-phosphate-linked monosaccharides and oligosaccharides and an acceptor substrate includes either oligosaccharides or proteins and ceramide for glycoproteins (Stanley et al., 2009).

During glycosylation, glycosyltransferase enzymes use acceptor substrates such as oligosaccharides, monosaccharides, polypeptides, lipids and other organic molecules to initiate the synthesis of glycoconjugates. (Pinho & Reis, 2015) (Stanley et al., 2009) The glycosyltransferase enzymes act sequentially to produce acceptor substrates that will be used in other subsequent reactions and these enzymes primarily plays role in the elongation of glycan chains that results in a linear or a branched polymer that has monosaccharides linked to one another. (Stanley et al., 2009) There are glycosidases that are involved also in the biosynthesis of glycans. These enzymes generate intermediates that can be acted by glycosyltransferase by removing monosaccharides and this is relevant process especially in the formation of N-glycans. For instance, glycosidases along with mannosidases can sequentially trim the nascent glycoprotein glycan so that glycosyltransferases can modify them to form a more complex and hybrid-type chains. There are

several glycan-modifying enzymes that utilizes variety of different donors such as sulfotransferases, phosphotransferases, O-acetyl-transferases and O-methyltransferases. (Stanley et al., 2009) (Paulson & Colley, 1989) Overall, a clear understanding of functional roles of glycosyltransferase is required to better understand how and what altered glycan changes could result in development and progression of cancer states.

3.2 Glycobiology in cancer

Glycobiology has recently gained an increased attention in cancer researches to understand the mechanisms of cancer and set up and develop therapeutic strategies and diagnostics. Glycosylation as described in section 3.1.2, is a common form of posttranslational modification (PTM) and this process plays an important role as a regulatory mechanism controlling various physio-pathological processes in a cell. (Pinho & Reis, 2015) Different types of glycoconjugates generated through glycosylation in a cell plays roles in cancer cell processes as well as the tumour microenvironment that leads to cancer progression. The compositional changes of glycans that have been added to either glycoproteins or glycolipids can aid various stages of cancer progression and this has been illustrated in figure 3.4. (Hakomori, 2002) (Munkley & Elliott, 2016) The biochemical mechanisms of altered glycan production remain poorly understood and it is predicted that the alteration of glycan structures occurs due to several reasons including, genetic mutations, changes in epigenetics and abnormal regulation of glycosyltransferases and chaperone genes. (Munkley & Elliott, 2016)

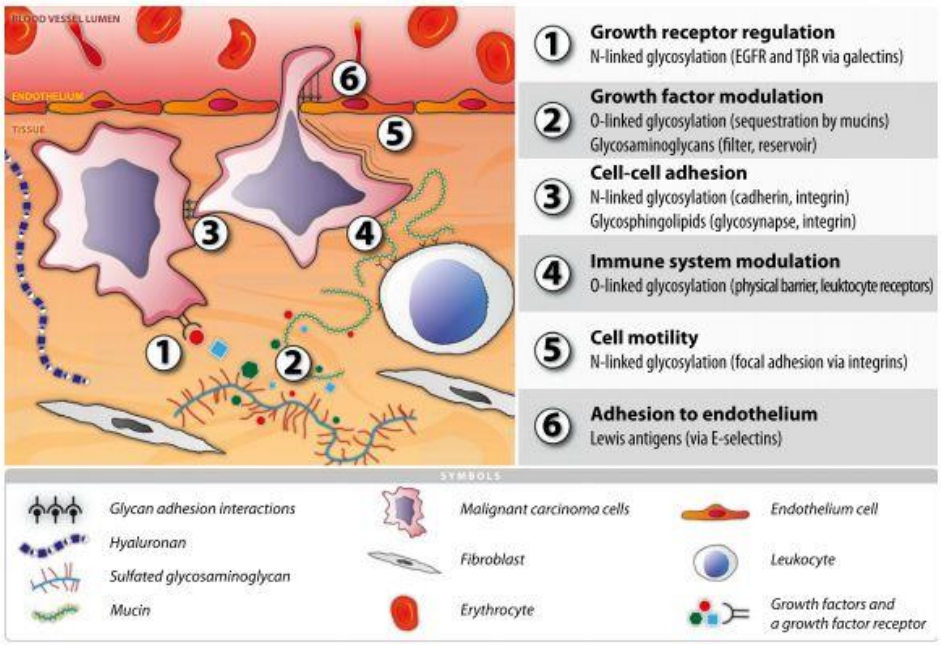


Figure 3.4 Glycosylation processes in carcinogenesis. Six important steps involved in the metastasis of carcinoma cells are illustrated. Initially, N-glycosylation influences the growth receptor and the concentration of growth factors increases. The N-linked glycosylation mediates cell to cell adhesion and O-glycosylated mucins act on a specific leukocyte and initiates immune system response targeting the malignant cells. The integrin functionalities are regulated by N-linked glycosylation to enhance motility of transformed cells. Finally, via binding of Lewis antigens by endothelial selectins, endothelium adhesion is mediated (Potapenko et al., 2010).

The altered glycosylation is closely linked to the progression of cancer in several ways (figure 3.4). The glycosylation change affects and modifies the cell surface carbohydrate structures that play significant roles in cell to cell interaction, signalling and adhesion properties that are essential for a tumour cell metastasis along with tumour cell interaction with the immune system. (Potapenko et al., 2010) Typically, malignancy introduced due to aberrant changes in glycosylation results in the creation of novel structures, precursor accumulation, persistence of truncated structures and a decreased expression of certain structures in a cell. (Hakomori, 2002)

Through aberrant glycosylation, the molecular heterogeneity and functional diversity can be introduced. Aberrant glycosylation is protein specific, site specific and cell specific and this specificity of glycosylation process is largely dependent on factors such as incomplete synthesis and neo-synthesis processes. The incomplete synthesis process usually occurs in early stages of cancer and this process involves biosynthesis of truncated structures due to malfunctioning of normal complex glycan synthesis process in the epithelial cells. In contrast, neo-synthesis processes are frequent in advanced stages of cancer whereby cancer associated genes are induced and they are involved in the direct expression of carbohydrates. (Pinho & Reis, 2015) It has been seen that aberrant glycosylation that leads to altered glycan structures in cancer cells can be attributed to dynamic expression changes of glycosyltransferase at the transcriptional level, tertiary conformation of the peptide backbone and localization of appropriate glycosyltransferases in the Golgi apparatus. (Pinho & Reis, 2015) (Reis et al., 2010)

With a reference to breast cancer, the changes in PTM of proteins that occur during neoplastic transformation affects breast cancer and with this, an understanding of PTM changes that contribute to oncogenic cancer progression would significantly aid the development of anticancer agents that prevents these PTM. (Krueger & Srivastava, 2006) There are four main glycan groups that are involved in target attachment namely, N-linked glycans, O-linked glycans, glycosaminoglycans (GAGs) and polysaccharides. (Potapenko et al., 2010) The N-linked, O-linked glycans and GAGs are attached to the polypeptide chains. The glycosphingolipids from GAGs form most glycolipids in a cell. The N-glycans are involved in growth, differentiation, adhesion and metastasis which is important for tumour cell progression. The O-glycans are involved in adhesion and immune response modulation. (Pinho & Reis, 2015) (Meany & Chan, 2011)

Sialylation and fucosylation are one of the most widely occurring cancer related glycosylation changes along with N- and O-linked glycan branching. The α2,6- and α2,3-linked sialylation has been identified to be most closely associated with cancer, involved in various cellular functions like cell recognition, adhesion and signalling. This process is typically catalysed by ST3Gal-I, sialyltransferase enzymes. There are several subgroups of sialyltransferase enzymes and these are type II transmembrane glycoproteins with a short amino (NH_2) cytoplasmic tail that resides in the lumen of Golgi. (Pinho & Reis, 2015) (Brockhausen, 2006) Fucosylated glycans can be

synthesized by fucosyltransferases (encoded by FUT). In fucosylation, α1,6-fucouse is added to the GlcNAc residue of N-glycans catalysed by FUT8 enzyme and overexpression of this type of enzyme typically influences breast cancer. The increase in dimerization and phosphorylation in breast cancer due to increased fucosylation can result in the increase of epidermal growth factor receptor (EGFR) mediated signalling and this plays a significant role in tumour cell growth and development of malignancy. (Pinho & Reis, 2015) (Kölbl, Andergassen, & Jeschke, 2015)

The glycosyltransferases (section 3.1.3) have been used to characterize glycosylation pathways and the understanding of the roles of glycosyltransferases in glycosylation enzymatic processes and a thorough study of glycosyltransferase expression at the transcriptional level therefore, have become a critical step to developing a biomarker and anticancer agents for diseases like breast cancer. Additionally, an increasing amount of glycomics, glycoproteomics and transcriptomics data will provide an exciting opportunity to discovery of novel targets and strategies for the diagnosis and treatment of diseases like breast cancer. In the next chapter, the breast cancer gene expression dataset and its pre-processing procedures followed for this project will be discussed.

3.3 References

1. Berninsone, P. M. (2005). Carbohydrates and glycosylation.
2. Brockhausen, I. (2006). Mucin-type O-glycans in human colon and breast cancer: glycodynamics and functions. EMBO reports, 7(6), 599-604.
3. Hakomori, S. (2002). Glycosylation defining cancer malignancy: new wine in an old bottle. Proceedings of the national academy of sciences, 99(16), 10231-10233.
4. Iha, H., & Yamada, M. (2013). Glycan profiling of adult T-cell leukemia (ATL) cells with the high resolution lectin microarrays T-Cell Leukemia-Characteristics, Treatment and Prevention: InTech.
5. Kailemia, M. J., Park, D., & Lebrilla, C. B. (2017). Glycans and glycoproteins as specific biomarkers for cancer. Analytical and bioanalytical chemistry, 409(2), 395-410.
6. Kölbl, A. C., Andergassen, U., & Jeschke, U. (2015). The role of glycosylation in breast cancer metastasis and cancer control. Frontiers in oncology, 5, 219.
7. Krueger, K. E., & Srivastava, S. (2006). Posttranslational protein modifications current implications for cancer detection, prevention, and therapeutics. Molecular & Cellular Proteomics, 5(10), 1799-1810.

8. Li, M., Song, L., & Qin, X. (2010). Glycan changes: cancer metastasis and anti-cancer vaccines. Journal of biosciences, 35(4), 665-673.

9. Meany, D. L., & Chan, D. W. (2011). Aberrant glycosylation associated with enzymes as cancer biomarkers. Clinical proteomics, 8(1), 7.

10. Munkley, J., & Elliott, D. J. (2016). Hallmarks of glycosylation in cancer. Oncotarget, 7(23), 35478.

11. Paulson, J. C., & Colley, K. J. (1989). Glycosyltransferases. Structure, localization, and control of cell type-specific glycosylation. Journal of Biological Chemistry, 264(30), 17615-17618.

12. Pinho, S. S., & Reis, C. A. (2015). Glycosylation in cancer: mechanisms and clinical implications. Nature Reviews Cancer, 15(9), 540.

13. Potapenko, I. O., Haakensen, V. D., Lüders, T., Helland, Å., Bukholm, I., Sørlie, T., . . . Børresen-Dale, A.-L. (2010). Glycan gene expression signatures in normal and malignant breast tissue; possible role in diagnosis and progression. Molecular oncology, 4(2), 98-118.

14. Reis, C. A., Osorio, H., Silva, L., Gomes, C., & David, L. (2010). Alterations in glycosylation as biomarkers for cancer detection. Journal of clinical pathology, 63(4), 322-329.

15. Stanley, P., Schachter, H., & Taniguchi, N. (2009). Essentials of glycobiology. Varki, A.

4 GENE EXPRESSION DATA

4.1 Gene expression data overview

The microarray gene expression dataset used in this work was retrieved from the CMap website (available at: http://www.broadinstitute.org/cmap/). In this dataset, gene expression was measured in the MCF-7 human breast cancer cell line, in two types of samples. The first samples were treated with several FDA-approved compounds, and the second type were untreated control samples. The dataset was loaded in RStudio, a statistical programming development environment, and several indispensable cleaning and pre-processing steps were followed before constructing a complete numerical gene expression matrix for subsequent analyses. The complete gene expression matrix was annotated using necessary annotation libraries and finally, the known GT gene list was matched against annotated genes to construct separate expression matrices with GT genes. The expression matrices were constructed and formatted to perform differential gene expression analysis and WGCNA.

4.2 The Connectivity Map

The advancement of human genome sequencing technologies, such as NGS techniques, has led to an explosion of new insights into the genetic basis of disease. Through sequencing analysis, disease-associated genes are identifiable, but their cellular functions remain unclear. From the chemical biology and drug discovery perspectives, new drugs were designed and developed with the use of advanced screening methodologies. This has led to the systemic establishment of chemical libraries and clinical therapeutics have entered the pharmaceutical markets (Lamb et al., 2006). However, it remains unclear what methods must be deployed to systematically determine the cellular effects of these novel therapeutics to potentially reduce side effects that limit clinical use.

The CMap project was initiated in 2006 at the Broad Institute, a biomedical and genomic research centre located in Cambridge, Massachusetts, USA, by (Lamb et al., 2006). Lamb et al. (2006) have hypothesized that the creation of a comprehensive genome-scale library of cellular signatures that catalogues transcriptional responses to chemical, genetic, and disease-related perturbations

would potentially solve the above-mentioned issues. It is predicted that previously unknown connections between several proteins, or between compounds with similar functions but different structures, could be unveiled through the examination of gene signatures with high similarity. Ultimately, the CMap project aimed to establish systemic relations among diseases, physiological processes, and the action of therapeutic molecules. This has led to the creation of a large public database of genome-wide transcriptional expression data with pattern-matching tools to detect similarities among gene signatures. Lamb et al. have named this tool the Connectivity Map (Lamb et al., 2006). The CMap has a web interface that could serve as a functional look-up table for the genome and, in connection with its interface, the database provides actual transcriptional expression profiles used to identify their gene signatures (Lamb et al., 2006). It is expected that CMap tools could aid researchers in the field of computational biology and chemical biology to better understand drug candidates and genes in the future.

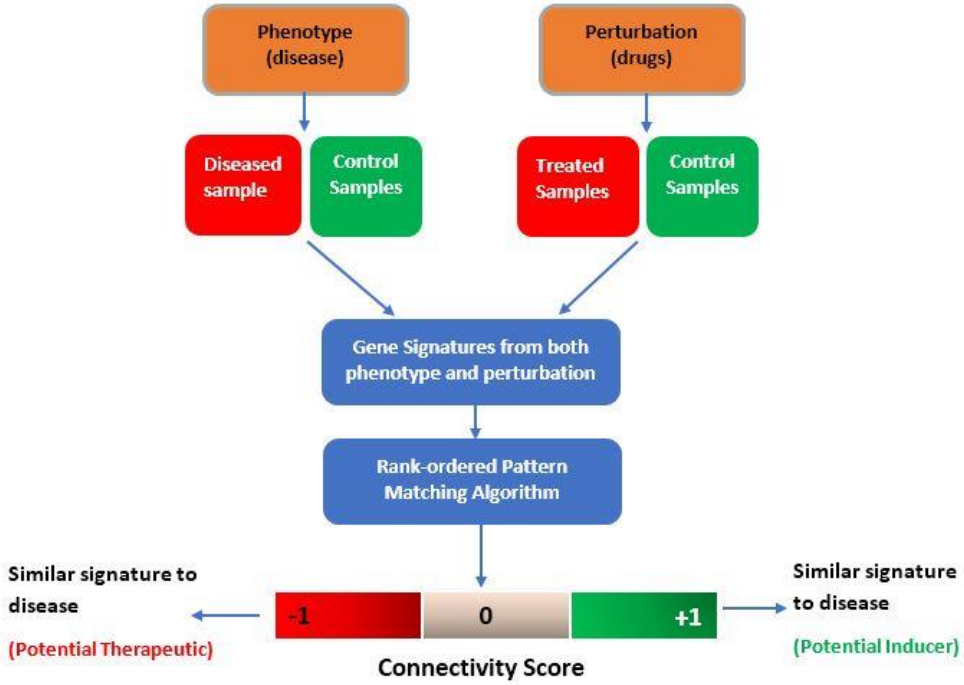


Figure 4.1 Schematic view of the Connectivity Map (CMap) concept and process. CMap provides a comparison method to measure similarities between a set of phenotype gene signatures and reference profiles from cell lines treated with FDA-approved compounds, and corresponding controls. The rank-ordered pattern-matching algorithm calculates a connectivity score between each pair of gene expression sets. The scores can be interpreted to identify potential therapeutic or inducer candidates.

4.2.1 Data

The current version of CMap, which includes L1000 assay data, was released on September 12, 2017, and is available on the cloud-based CLUE software platform together with online gene expression analysis tools (available at: https://clue.io/). The newest library contains over 1.5 million gene expression profiles from 5,000 therapeutic compounds and 3,000 genetic reagents. The earlier Affymetrix microarray-based CMap data (build 02) are publicly available via Gene Expression Omnibus (GEO) or can be manually downloaded in batch files from the Broad Institute website (available at: https://www.broadinstitute.org/). CMap build 02 was used in this work and the dataset contains more than 7,000 gene expression profiles representing 1,309 FDA-approved therapeutic compounds measured in five different cell lines. There are many applications of CMap datasets and the next section provides a review of the literature relating to these applications.

4.2.1.1 Cell lines

A cell line is developed from a single cell cultured in a given medium. It contains uniform genetic material and is widely used to perform biological studies such as drug metabolism testing, cytotoxicity, and gene function analyses. A usual caveat when using cell lines to perform biological research is that one must not have blind faith in the output derived from cell lines, as these may not always replicate primary cells accurately. However, cell lines have several advantages when performing scientific studies, such as cost-effectiveness and convenience (Lee, Kuo, Whitmore, & Sklar, 2000). Lamb et al. generated transcriptional expression profiles based on four different cell lines: these are listed in table 4.1, along with the number of experiments performed and the number of different drugs used to treat the respective cell lines (Lamb et al., 2006).

Table 4.1 Cell lines used in the Connectivity Map database

Cell line	Type	Experiments	No. of drugs used
MCF-7	Breast cancer	3095	1294
PC3	Prostate cancer	1741	1182
HL60	Leukaemia cell	1229	1078

SKMEL5	Melanoma	17	17

Lamb et al. generated most of their data in the MCF-7 breast cancer epithelial cell line as it is by far the most standard reference cell line in laboratories around the world and has been molecularly characterized most extensively (Lamb et al., 2006).

4.2.1.2 Perturbagens

Perturbagens are biological substances that disrupt intracellular processes. The very first CMap database (build 01) studied the effects of 164 distinct small molecule perturbagens on different cell lines (Lamb, 2007). For the work presented here, the CMap build 02 database was used. According to Lamb et al., the perturbagens used in their work range from FDA-approved drugs to non-drug bioactive compounds (Lamb et al., 2006). To identify any compounds sharing a molecular signature, compounds with shared targets were considered. Compounds with the same clinical indications were included in this profiling to infer connections between therapeutic classes, although the compounds' chemical structures and molecular mechanisms of action were different. A detailed description of specific compounds used in this research work is presented in the section 4.4.

4.2.1.3 Concentration and duration of treatment

Lamb et al. performed expression profiling using a range of treatment concentrations, including concentrations that were reported to be effective in different cell lines, to explore the effect of treatment dose in expression profiling (Lamb et al., 2006). The duration of treatment was uniform across all compounds. Expression levels were measured 6 h after the addition of compounds as the goal was to observe gene signatures related to direct mechanisms of action. For this research work, only cell samples treated with the highest drug concentration were selected for subsequent analyses.

4.2.1.4 Gene expression profiling methods

The following section describes how Lamb et al. generated their gene expression profiles (Lamb et al., 2006). This collection of gene expression profiles from different treatment and control samples forms the CMap database. Expression profiling was done in batches of either 6-well dishes (known as the development batch) with four compound treatments and one corresponding control, or 96-well dishes (known as the production batch) with 42 compound treatments and six controls. Cells were treated 24 h after plating. Full details of the CMap drug treatment dataset (build 02) are available at http://www.broad.mit.edu/cmap as "cmap_instance.txt".

TRIzol (Invitrogen) was used to isolate total RNA. Total isolated RNA samples were hybridized to microarray for complementary RNA (cRNA) synthesis, and scanning was done using Affymetrix GeneChip products. Total RNA extracted from development batches was processed using the HG-U133A array, while total RNA from production batches was processed using the HT_HG-U133A array. An HT scanner from Affymetrix (supporting information available at: https://www.affymetrix.com) was used to scan high-throughput arrays (HTA) and produce image files that contain numerical expression levels across genes. Faulty scans that did not meet the required basic data quality were discarded (Lamb, 2007). The dataset is available online in the CMap database (http://www.broad.mit.edu/cmap).

4.3 Applications of CMap

Since the advent of transcriptional gene expression profiling methods, many successful and radical biomedical science discoveries have been made. Using this technology, Lamb et al. initiated the CMap project, and its tools have successfully provided data-driven and systemic approaches to unveiling connections between genes, chemicals, and biological conditions such as diseases (Lamb et al., 2006). Since its introduction in 2006, many applications of CMap have shown promising results. The CMap tools and the database provided by the web interface have especially contributed to drug discovery and development, specifically in identifying new indications for existing drugs and elucidating the mode of action of novel chemicals (K. Wang et al., 2015).

Drug repurposing is one of the most popular applications of CMap. Drugs available on the market today have more than one effect or side effect, and bioactive compounds are prescribed for several

different indications (K. Wang et al., 2015). Pharmaceutical industries are continuously looking to reduce the cost, risk, and time-to-market associated with their drug products by employing drug repurposing strategies, which aim to identify new indications for existing drugs.

In addition to drug repurposing, CMap also has applications in lead discovery to identify novel active compounds for pharmaceutical pipelines across various disease areas. Traditionally, target-based drug discovery prioritized identification of putative therapeutic drug targets before compound screening. However, CMap has employed a phenotype-based approach that does not require a deep understanding of cellular regulatory networks and regulatory module functions. Researchers could simply query CMap using the gene signature of a diseased tissue sample to identify therapeutic molecules that could modulate this disease state (K. Wang et al., 2015). The use of CMap in drug repurposing and lead discovery has yielded significant results and it is expected that the use of established tools such as CMap for drug research will continue to increase.

In the next few paragraphs, the impressive applications of CMap in drug repurposing, along with several other applications including lead discovery, will be discussed.

4.3.1 Drug repurposing via machine learning

Napolitano et al. developed a machine learning computational tool to predict drug repurposing (Napolitano et al., 2013). According to their studies, many traditional computational approaches to drug repurposing focused only on changes in transcriptional expression levels in response to drug treatment, or on drug-disease relationships. Specifically, correlations between drug-associated and disease-associated expression signatures were examined. Although this type of approach was able to identify and validate some novel therapeutic indications, the noisy and complex nature of gene expression data and the limited availability of genomic data for many diseases slowed down drug repurposing studies. For instance, publicly available gene expression data were generated from patients that had already been treated with other drugs and merging several information levels together makes the drug repurposing process less reliable and robust.

To overcome these constraints, the authors focused on enhancing the predictability of existing computational tools with machine learning classification algorithms by incorporating drug characteristics and not considering data concerning the disease (Napolitano et al., 2013). In this

way, they were able to present mismatches between known and predicted drug classification. The mismatches were interpreted as potential alternative therapeutic indications.

The traditional supervised machine learning classification algorithm Support Vector Machine (SVM) was used to build the classifier. The authors predicted drug repurposing based on the similarity of chemical structures, the proximity of their targets within the protein-protein interaction network, and finally, gene expression correlation patterns after treatment with drugs. The gene expression profiles used to train the machine learning classifier were derived from the CMap database (Napolitano et al., 2013).

The classifier achieved an accuracy of 78% and based on this, the authors re-interpreted the top misclassifications as re-classifications after several statistical procedures (Napolitano et al., 2013). The approach followed by this study maximized re-classification efficacy towards the repurposing of known drugs and novel drug discovery (Napolitano et al., 2013). The following flowchart summarizes the analysis pipeline followed by the authors.

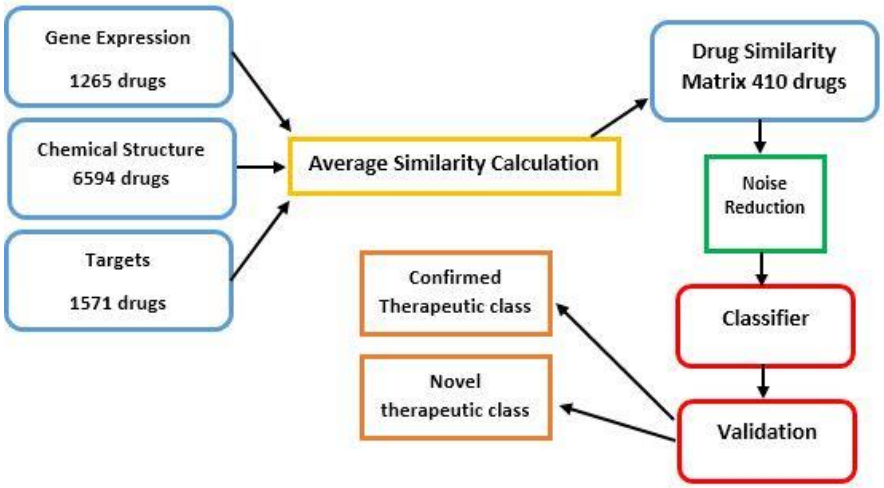


Figure 4.2 Flowchart of the machine learning classifier development process followed by Wang et al. (K. Wang et al., 2015). Blue boxes indicate data while other colours indicate steps in the development process.

4.3.2 Drug functional similarity analysis

Cha et al. aimed to identify the functional similarity of different drugs by comparing the transcriptional profiles of CMap samples in response to drug treatment (Cha, Kim, Oh, Shin, & Yi, 2014). CMap data are especially useful in unveiling the unknown characteristics of small compounds if multiple features are considered. The authors diversified features of comparison in gene expression by combining signatures of differentially expressed genes (DEGs) and differentially co-expressed modules (DCMs), and this integration provided more robust and accurate methods of searching for similar drugs based on expression profiles (Cha et al., 2014). Co-expressed gene modules, in addition to DEGs, can be very useful features, and several studies have identified certain DCMs involved in cancer development, even though differences in expression levels were small (Lai, Wu, Chen, & Zhao, 2004; Carter, Brechbühler, Griffin, & Bond, 2004; Kostka & Spang, 2004).

Cha et al. modified the original linear modelling test that employs a t-test to normalise different experimental samples. Significant DCMs were identified using maximal cliques among the co-expressed modules. Similarity search methods were evaluated by producing a receiver operating characteristics (ROC) curve, and an overall performance score of 0.99 was calculated by measuring the area under the curve (AUC). Additionally, the authors constructed a drug-drug network using similarity search methods to identify drugs having similar gene expression responses. In this way, similarity search methods could provide insights into drug repurposing using CMap expression profiles. The biggest limitation of this study was the small number of drug response samples, as the robust and reliable identification of DCMs requires sufficient sample size. (Cha et al., 2014)

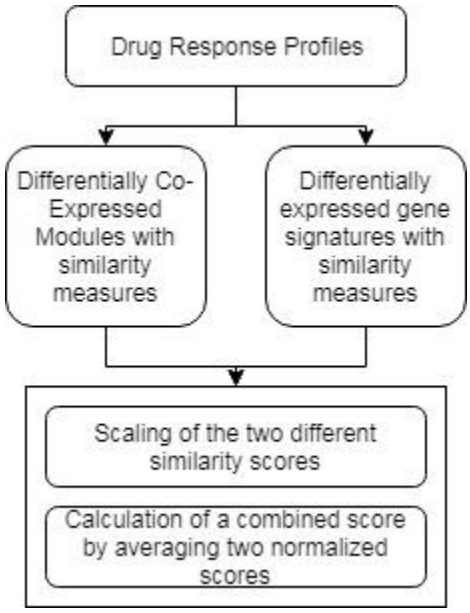


Figure 4.3 Overall procedures adopted by Cha et al. to derive their similarity search methodology (Cha et al., 2014). Differentially co-expressed gene module (DCM) signatures and differentially expressed gene (DEG) signatures were identified, and combined scores for the two different signatures were calculated by considering the false discovery rate (FDR).

4.3.3 Drug safety evaluation

Recently, the popularity of using CMap in drug repurposing research has increased, but much less attention is given to the potential use of CMap in evaluating drug safety and side effects. Adverse drug reactions (ADRs), also known as side effects, are important factors that pharmaceutical industries consider when developing new drugs. Wang et al. recently developed a drug safety evaluation method using the CMap database to eliminate any potential risks of drugs in the preclinical stage of development (Ashburn & Thor, 2004; K. Wang et al., 2015). The authors computationally designed a risk score model that predicts drug safety risks based on drug-induced gene-gene expression profiles from the CMap database (Ashburn & Thor, 2004). Their results showed that the transcriptional expression profiles of drugs that induce ADRs were positively correlated with expression profiles in CMap, allowing the authors to effectively identify drugs with safety risks. Once again, this study showed how integration of the CMap database and computational modelling and analysis can be very useful in the preclinical stage of drug development (Ashburn & Thor, 2004).

4.3.4 Lead molecule discovery

Lead molecule discovery is another well-known application of CMap as discussed in section 4.3. Wang et al. performed a meta-analysis of two sets of published microarray data (from cancerous and non-cancerous samples) to obtain a gene signature with 343 DEGs specific to lung adenocarcinoma (G. Wang et al., 2011). The DEGs were submitted to the CMap analysis interface to identify drugs that show negative expression correlation with the gene signature obtained in their own analysis. The authors identified several compounds that showed negative correlation, and selected 17-AAG, an inhibitor of heat shock protein 90 (HSP90), as a lead molecule that suppressed growth of lung adenocarcinoma cells *in vitro* (Kaur & Dufour, 2012).

In conclusion, the applications of CMap in drug development programs have been highlighted in the previous paragraphs. Although CMap has its own limitations, it provides valuable data and systematic approaches that can be adopted in drug discovery programs to generate testable hypotheses. It is expected that the use of CMap in drug discovery research, including drug repurposing and lead molecule discovery, will continue to increase in the future (K. Wang et al., 2015).

4.4 Data preparation

This section describes how expression data were initially prepared for this research work. Data from gene expression in response to compound treatment were downloaded from the CMap database available online. A total number of 7,056 (.CEL) raw expression files were downloaded together with the instance (description) file.

4.4.1 Data cleaning

Data cleaning and filtering form essential part of microarray data analysis in order to select only necessary features and discard unnecessary ones. The CMap instance file was loaded in RStudio running R version 3.3.3 with necessary data manipulation packages such as “dplyr”, “plyr”, and “stringr”. Only treated MCF-7 cell samples were selected for subsequent preparation procedures. There were three types of Affymetrix chips used to generate expression profiles, namely HG-U133A, HT-HG-U133A, and HT_HG-U133_EA. HT_HG-U133A_EA is an early access version

of the Affymetrix chip and, due to the small number of samples included, this chip type was discarded. In total, 2,911 samples were filtered, of which 2,740 samples were from the HT_HG-U133A chip and 171 samples from the HG-U133A chip.

4.4.2 Biological replicates and drug information

Typically, high-throughput microarray experiments for genetic studies are costly and time-consuming. The production of arrays for experiments can be slow and supply may be limited. Due to these reasons, the manufacturers of microarrays sometimes minimize the importance of replication in scientific experiments and may underestimate the significance of including biological replicates in their studies (Lee et al., 2000). Biological replicates are a set of multiple samples of the same tissue measured across multiple treatments. It is expected that these samples respond very similarly to treatment. Biological replicates are necessary in microarray data analysis as they are used to test variability between samples and confirm that the difference between treated samples is truly due to biological variation. There are several advantages to including sufficient biological replicates in microarray experiments. First, replicates improve the measurement of variation and the statistical significance of tests performed. This is especially important in differential gene expression analysis. Second, replicates can be used to detect outliers. Finally, precision of gene expression is improved by averaging across replicates (Newton, Kendziorski, Richmond, Blattner, & Tsui, 2001).

In this study, compound-treated samples with at least five biological replicates were selected. Control samples (untreated MCF-7 cells) also comprised five biological replicates. In total, 29 drugs fulfilled this criterion, each with 10 samples: five treated samples and five control samples. Five treated samples were randomly selected from each drug to reduce bias by the number of experiments. For example, there were 181 replicates treated with Trichostatin A, from which five replicates were randomly selected. Finally, all the samples selected were treated with the highest drug concentration.

Information regarding the 29 drugs is available in table 1, appendix B. This information includes drug indications, target proteins, and chemical formulae, along with PubChem compound identifiers (CIDs). PubChem is a chemical molecule database that contains compound activities in biological assays. The system is maintained by the NCBI. PubChem CIDs for each drug have been

included in the table. These CIDs can be used to search for more detailed information about the compound including its structures, properties, medical information, pharmacology, biochemistry, and related literature.

4.5 Data pre-processing

Data pre-processing is indispensable in microarray data analysis. It is difficult to visually detect problems in high-throughput data such as microarray data, as the numerical gene expression matrices obtained are huge. This has led to the development of specific microarray data quality control procedures (Sánchez & de Villa, 2008). This section describes how the gene expression data were quality-controlled, pre-processed, and reformatted for subsequent statistical analyses.

The treated and control sample gene expression data (.CEL files) from 29 selected drugs were loaded in RStudio using necessary Bioconductor packages including the "affy" library. The total number of .CEL files loaded in RStudio for each chip type is shown in table 4.3. The shared controls were used for this dataset thus, there are more treated samples present in this study.

Table 4.3 Total number of expression files loaded in RStudio for pre-processing

Total: 203	**HG-U133A chip type**	**HT_HG-U133A chip type**
Treated	14	131
Control	12	46
Total	26	177

Several data quality checking procedures and normalizations were performed on the raw gene expression data. The next few sections describe data quality assessment and normalization procedures.

4.5.1 Data quality assessment

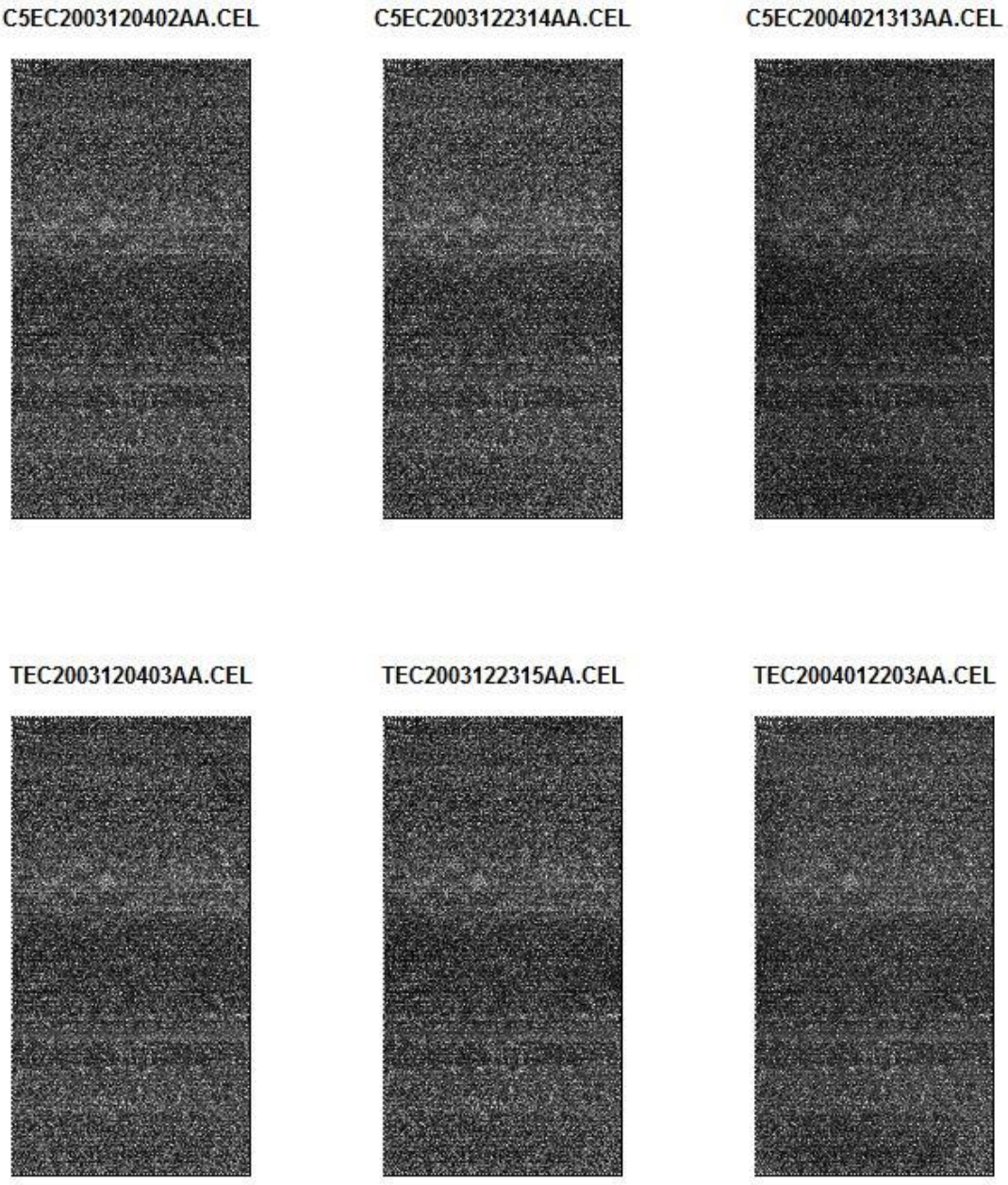
The main goal of data quality assessment is to remove any faults in the data and to determine if the data can be considered reliable. This research work utilized assessment tools such as 2D array image analysis and density plot analysis.

4.5.1.1 Image quality assessment

2D images of gene expression array positions can be used for spatial bias diagnostics. These images can be used to display spatial trends or biases that cannot be detected in the raw data. The raw probe intensity measurement can be used directly to plot images, and data quality can be assessed by examining non-random physical factors in the images, including, rings, shadows, lines, and strong variation in shade (Gillespie, Lei, Boys, Greenall, & Wilkinson, 2010).

The loaded gene expression samples for each chip type were used to generate 2D images and the first few 2D images from each chip type array are shown in figure 4.4.

(a)


C5EC2003120402AA.CEL
C5EC2003122314AA.CEL
C5EC2004021313AA.CEL
TEC2003120403AA.CEL
TEC2003122315AA.CEL
TEC2004012203AA.CEL

(b)

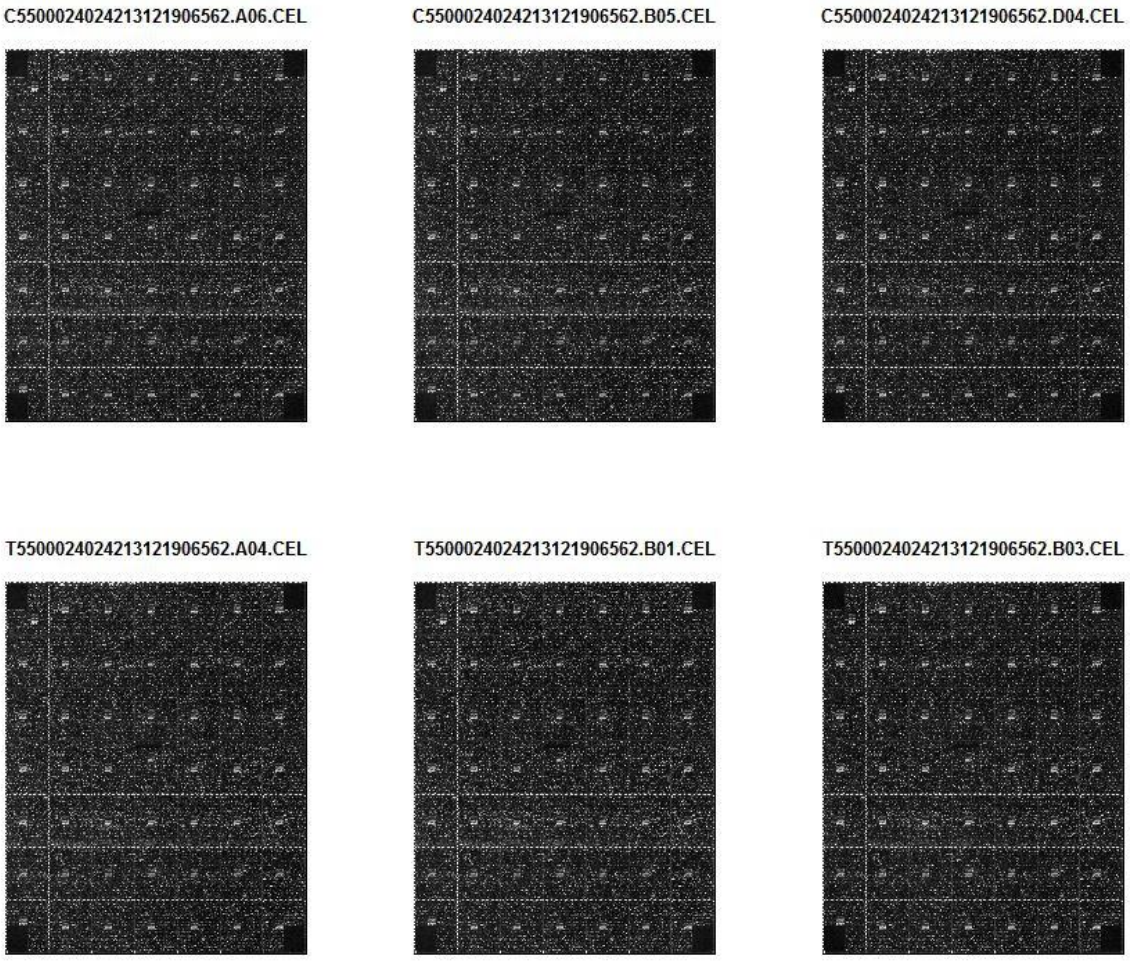


Figure 4.4 2D image plots of the mismatch and perfect match probe intensities of arrays from each chip type HG-U133A and HT_HG-U133A. Control sample arrays are represented by the letter "C", and treated samples are represented by the letter "T", along with the respective array number. (a) First six 2D images from the HG-U133A chip, representing control and treated samples. (b) First six 2D images from the HT_HG-U133A chip, representing control and treated samples. These samples were randomly selected from a total of 203 samples from each chip type. The rest of the sample 2D images were analysed separately.

From these 2D images, it is clear that datasets from each chip type did not appear to have any non-random structures, but maintained pattern consistency, and it can therefore be concluded that CMap data quality is high.

4.5.1.2 Density plot assessment

Evaluating density plots of probe intensities is another useful data quality assessment strategy. Density plots are generated using log-intensity distribution for each sample array, and the plots are super-posed on a single graph for better comparison between arrays. Density plots are especially useful as abnormal distributions can easily be detected when the different density plots are super-posed. For instance, differences in spread and position can be corrected by normalization, although significant multi-modality in the density plot distribution including any outlying observations could indicate that the data quality is questionable (Gillespie et al., 2010).

The density plots of log-intensity for all sample arrays from each chip type are shown in figure 4.5. These density plots were generated using RStudio.

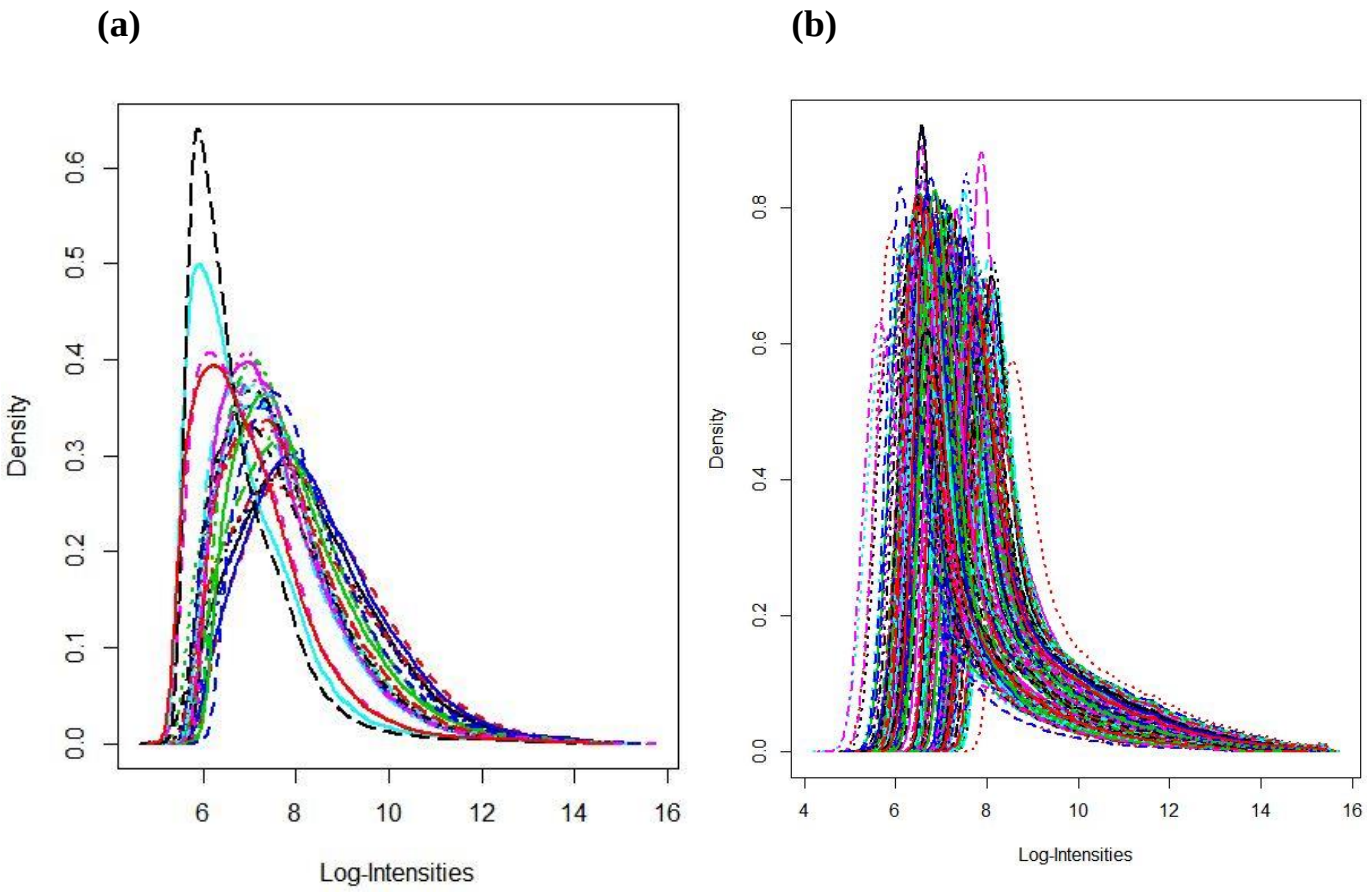


Figure 4.5 Density plots of log-intensity (x-axis) distribution of sample arrays from each chip type. Both graphs show treated and control samples together. (a) Density plot of 26 samples from the HG-U133A chip. (b) Density plot of 177 samples from the HT_HG-U133A chip.

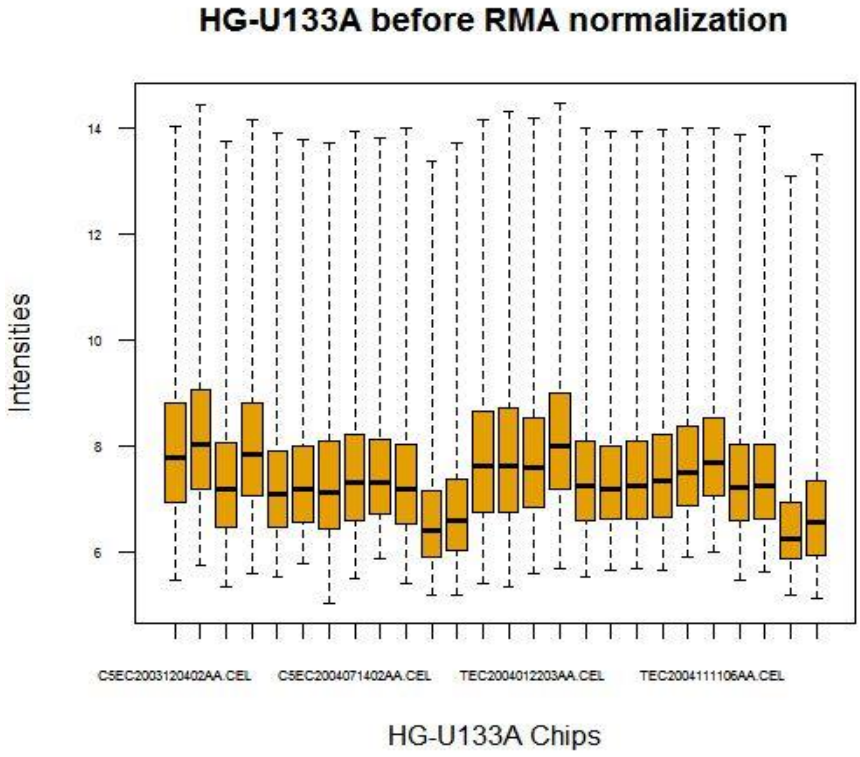
It is clearly visible from figure 4.5 that there are differences in the spread and positioning of the distributions, which can be corrected by normalization. It is also clear that there is no observed significant multi-modality in the distributions of density plots for each chip type and it can be concluded that CMap data are of high quality. Several data quality assessment tools can successfully be used to filter out problematic arrays and control the quality of the data.

4.5.2 Normalization

Normalization of microarray data must be performed to correct any technical and systemic biases caused by factors such as differential dye absorption and spatial heterogeneity in the arrays (Sánchez & de Villa, 2008). The next few paragraphs describe several normalization procedures performed on the CMap data including RMA and quantile normalization, along with boxplot visualization of arrays before and after different normalization steps.

Boxplot visualization of log-intensity distribution can be very useful to inspect different normalization steps and ensure that any technical and systemic biases have been removed from the data. Boxplots of samples from each chip type before normalization are shown in figure 4.6. Expression samples from each chip type were loaded in RStudio using the “affy” package.

(a)



(b)

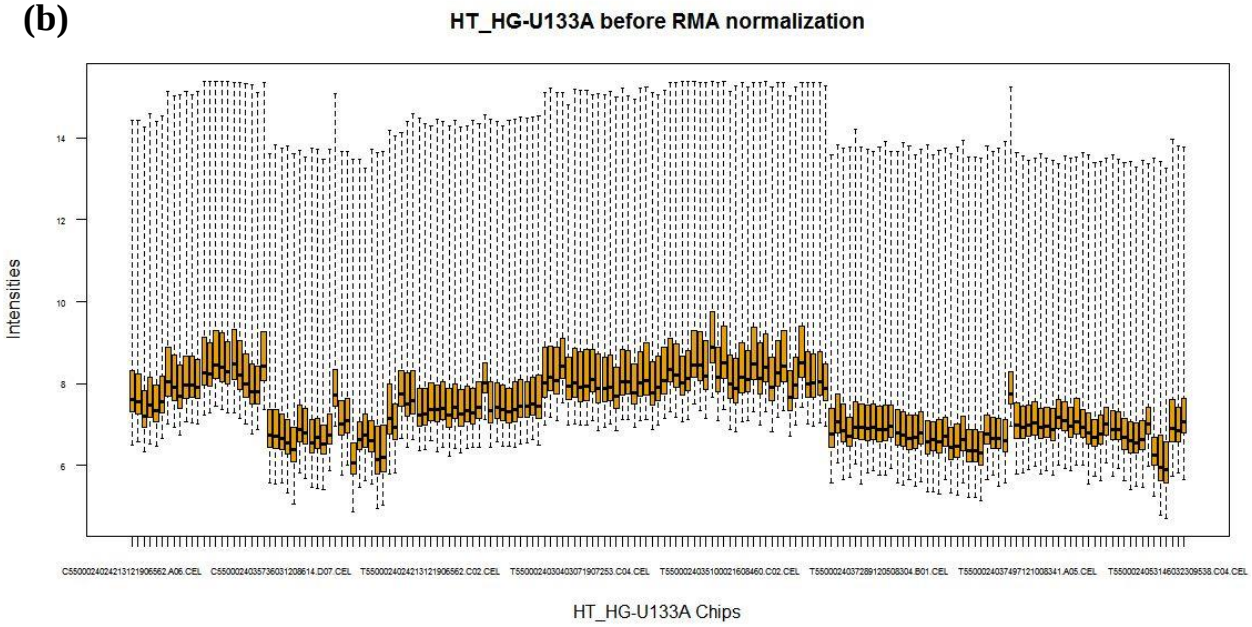


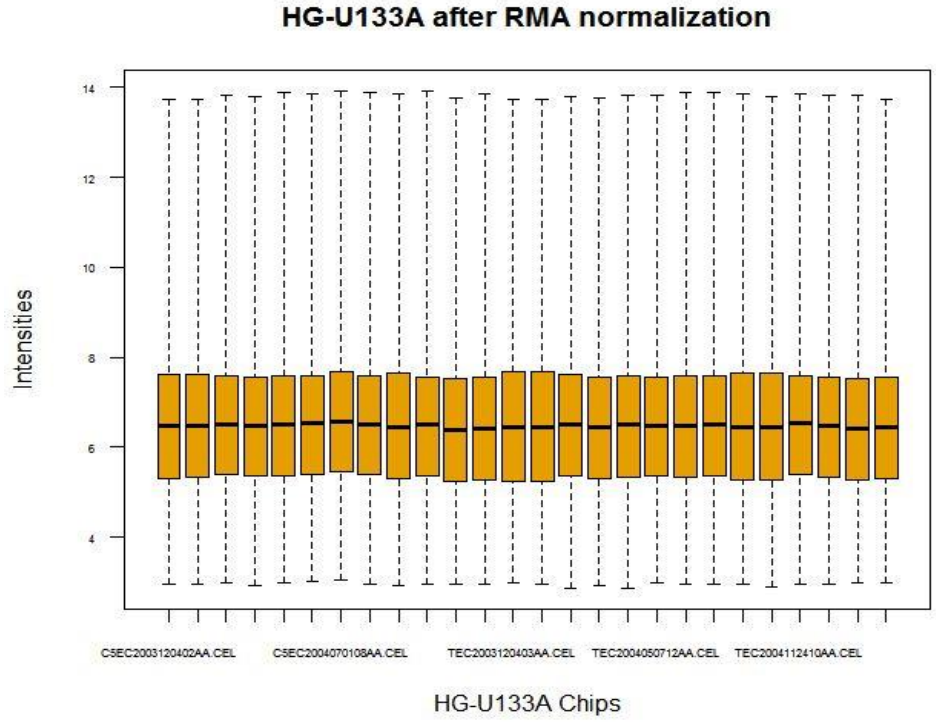
Figure 4.6 Boxplots of sample arrays from each chip type before RMA normalization. (a) HG-U133A chip (26 samples). (b) HT_HG-U133A chip (177 samples).

From the boxplots shown in figure 4.6, it is clear that the spread and position of sample probe intensities differ widely. This is due to technical and systemic biases that are introduced during microarray experiments and these must be removed to obtain more consistent probe intensities for subsequent statistical analyses.

4.5.2.1 RMA normalization

RMA normalization was designed for data obtained from oligonucleotide chips such as Affymetrix gene chips. RMA normalization combines model-based background correction, quantile normalization, and robust averaging of samples across hybridization sets at the probe level (Gillespie et al., 2010). RMA normalization returns log2-transformed expression values in the matrix. Expression matrices from each chip type were RMA-normalized using the "affy" package. Boxplots for each chip type obtained after RMA normalization are shown in figure 4.7. Spread and position were corrected by RMA normalization and more consistent probe intensities can be observed. After RMA normalization of each expression matrix, these can be merged into a single matrix for subsequent statistical analyses.

(a)



(b)

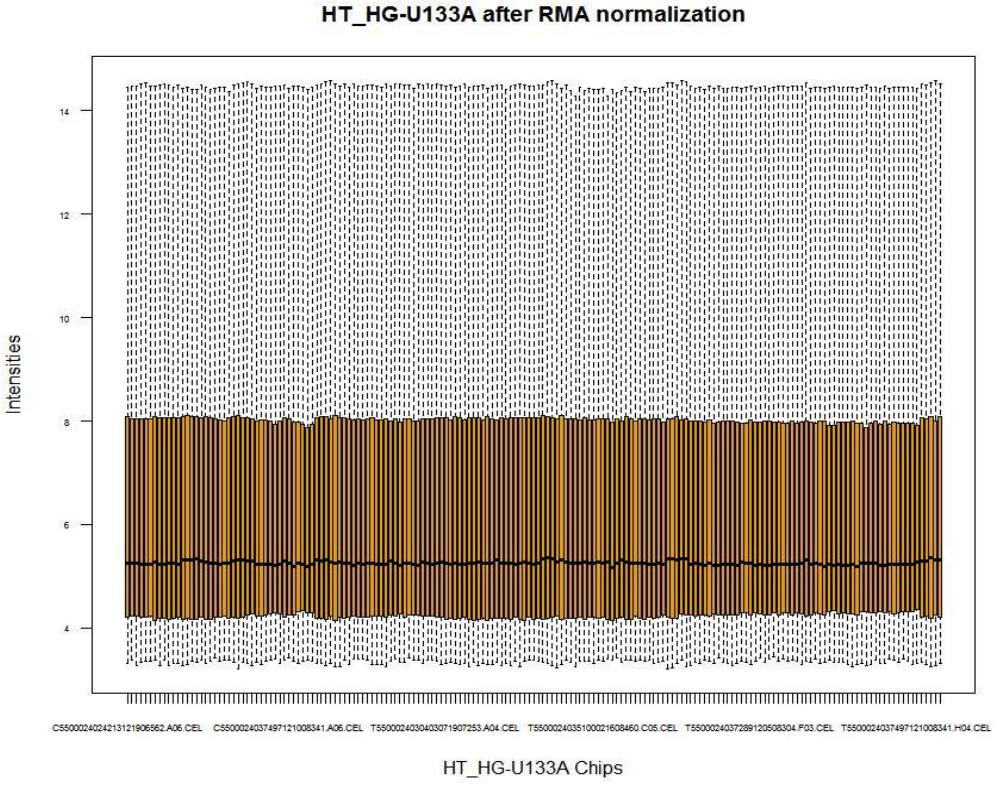


Figure 4.7 Boxplots of sample arrays from each chip type after RMA normalization. (a) HG-U133A chip (26 samples). (b) HT_HG-U133A chip (177 samples). More consistent probe intensities across samples can be observed after RMA normalization.

4.5.2.2 Quantile normalization

In many cases, the distribution of expression intensities between different arrays differs when arrays are combined, and it is necessary to normalize expression intensities for consistency between different arrays to obtain similar distributions (Sánchez & de Villa, 2008, Gillespie et al., 2010) This can be achieved by quantile normalization. Two different Affymetrix arrays, HG-U133A and HT_HG-U133A, were used in CMap expression profiling and the RMA-normalized expression matrices from each chip type were merged into a single expression matrix. At this point, rows in the expression matrix are labelled 'probe ID', and columns are labelled with the name of the treatment drug. One column is referred to as one drug-treated sample. Boxplots obtained from the merged expression matrix before and after quantile normalization are shown in figure 4.8. Quantile normalization of the merged expression matrix was performed using the "LIMMA" package in Bioconductor.

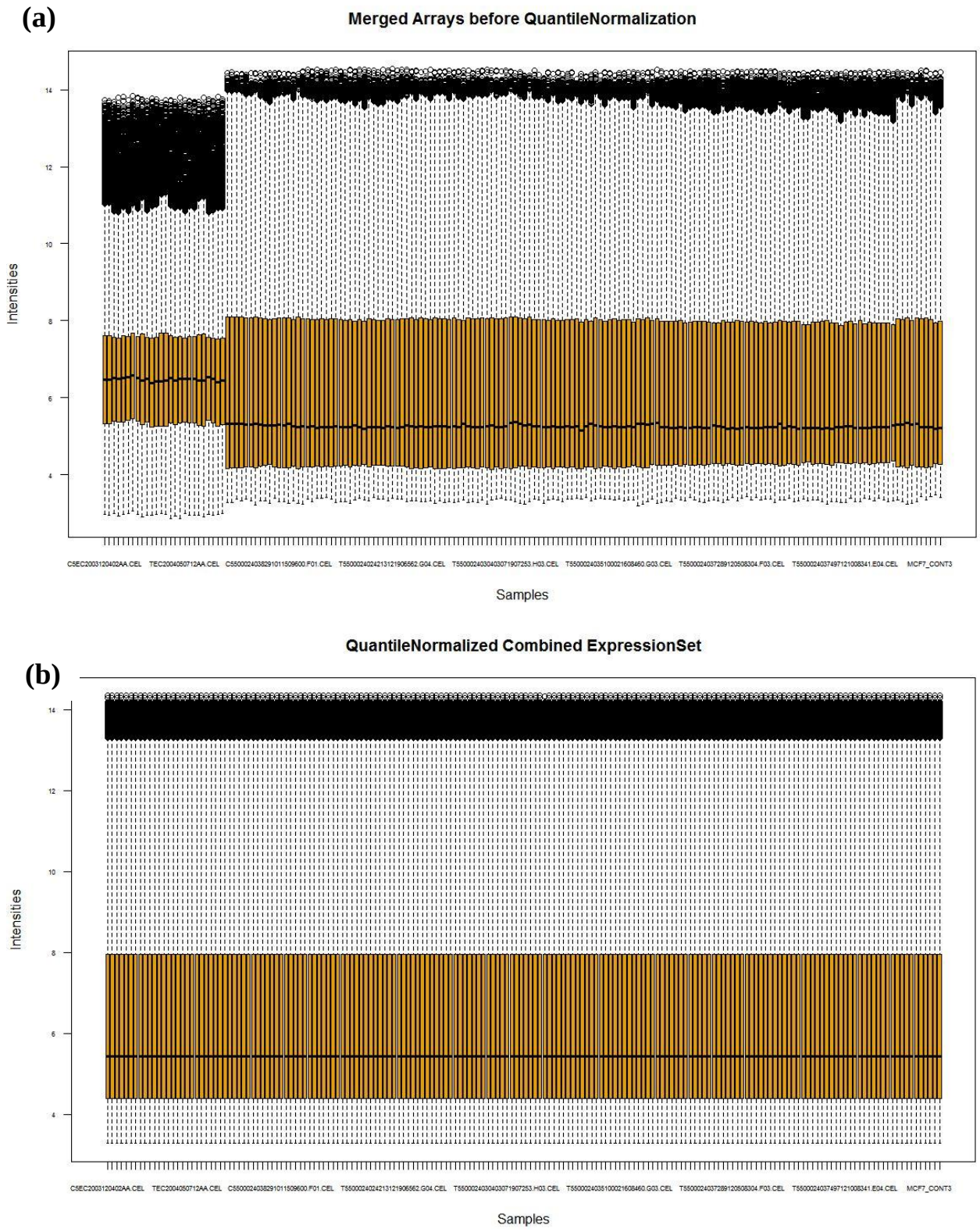


Figure 4.8 Boxplots of merged expression data frames before and after quantile normalization. (a) Before quantile normalization. (b) After quantile normalization.

The boxplot of the merged expression matrix displays a clear separation of spread, position, and distribution of intensities between the two chip types (figure 4.8 a). However, after quantile normalization (figure 4.8 b), consistency between arrays is achieved.

4.6 Batch effect detection

Batch effects are potential limitations of the use of transcriptomic datasets such as those found in the CMap database. Gene expression level changes due to treatment with a drug may be confounded by uncontrolled variables such as cell culture conditions and batch handling by the experimenter. Cell lines that are treated with the same drug in different batches may result in greater batch variation in expression levels and less so in drug action (Newton et al., 2001). Thus, batch effects must be detected before proceeding with statistical analyses, and if batch effects are present, these must be removed in controlled ways. The next few paragraphs describe how batch effects in the expression matrix were detected using PCA and how these were removed using the "ComBat" package in Bioconductor. PCA plots were used to display both expression matrices before and after batch effect removal.

4.6.1 Principle component analysis

PCA is a useful exploratory data analysis visualization tool that helps to reduce data dimensionality while still retaining much of variability in the data. PCA can also identify useful patterns in the data.

Drug-treated samples from the merged expression matrix in section 4.5.2.2 were grouped according to corresponding batch IDs. The samples were divided into two clusters after batch grouping, and the two clusters were labelled with the letter "A" and "B", respectively. Samples from batch cluster A were labelled with red and cluster B with black. PCA was performed using the "prcomp" function from the Stats package in RStudio. The first PCA plot for batch effects detection is shown in figure 4.9.

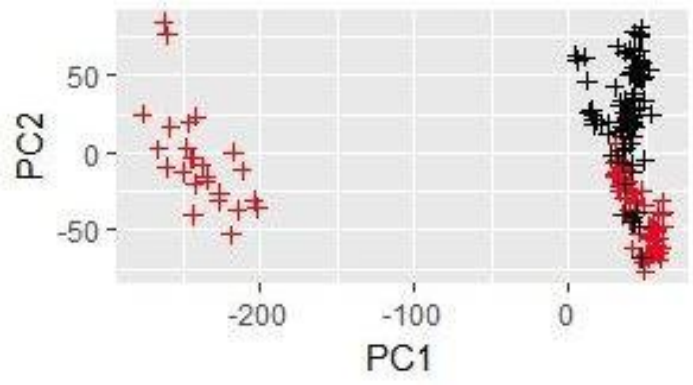


Figure 4.9 PCA plot of the merged expression matrix for detection of batch effects. Each cross in the plot represents a drug-treated expression sample. The red crosses represent batch cluster A samples and the black crosses represent batch cluster B samples, with x- and y-axes representing principal component scores.

PCA identifies linear combinations of variables such as probes and genes that explain the greatest variation in the data (Raychaudhuri et al., 1999). Typically, in gene expression analysis, PCA plots display how much influence each of the sample attributes selected from the linear combinations of variables has on the gene expression profile. Since the merged expression matrix was grouped by batch IDs and divided into two clusters using colour codes, if the plot separates expression samples into different clusters with corresponding colour, this probably indicates that the largest source of variation in the expression data is due to batch effects. In this case, figure 4.9 clearly displays a separation of samples into two batch clusters with some outliers. This indicates that the largest source of variation from the merged expression matrix was due to batch effects and these must be removed.

4.6.2 ComBat analysis

Batch effects were identified in the merged expression matrix as shown in figure 4.9. Batch effects were removed from the expression matrix using the "ComBat" function in the "sva" package in Bioconductor. "ComBat" allows users to adjust for batch effects in datasets in which the batch covariate is known, using methodology described by Leek et al. (Leek, Johnson, Parker, Jaffe, & Storey, 2012). It uses either parametric or non-parametric empirical Bayes frameworks to correct batch effects in datasets (Leek et al., 2012). After running ComBat analysis on a cleaned and

normalized expression matrix, the batch effect-corrected expression matrix was returned. This was visualized using PCA by once again selecting variables that explain the greatest variation in the data (figure 4.10).

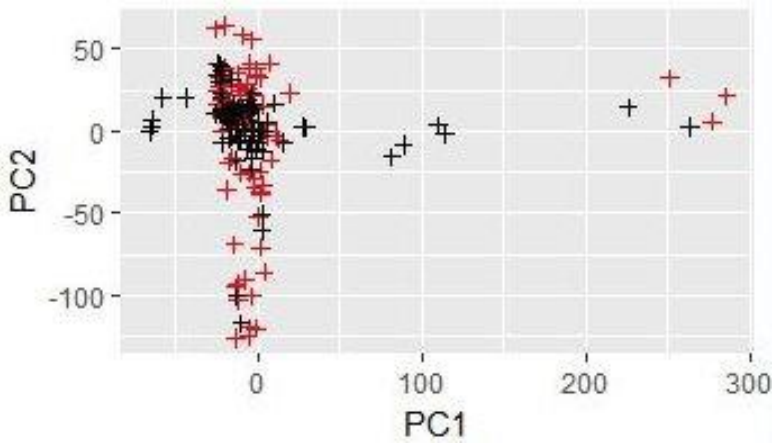


Figure 4.10 PCA plot of the expression matrix corrected for batch effects using ComBat analysis from the sva package in Bioconductor. The x- and y-axes represent principal component scores.

The PCA plot in figure 4.10 clearly displays clustering of two color-coded batch clusters with some outliers. This indicates that the largest source of variation is no longer related to batch effects. The batch effect-adjusted expression matrix returned can now be used for gene name annotation and subsequent statistical analyses. The pre-processing of expression data ends after batch effect correction.

4.7 Gene annotation and glycosyltransferase gene retrieval

The pre-processed numerical gene expression matrix contains 22,215 rows labelled with Affymetrix probe set IDs, and 203 columns corresponding to drug-treated samples. Each probe set ID must be annotated with its corresponding gene symbol. An annotation information file was generated using annotation libraries “hgu133a.db” and “hthgu133a.db” for chip types HG-U133A and HT_HG-U133A, respectively, in RStudio. When transforming probe set IDs to gene level data, some genes were represented by several probes set IDs, thus their mean intensities were used.

Since the U133A chip contains numerous genes with duplicated probe sets, this step was necessary and probes not matching gene names were discarded.

A set of 210 GT genes was retrieved from the carbohydrate-active enzymes (CAZy) database (http://www.cazy.org). The CAZy database contains detailed classification and related information about enzymes involved in carbohydrate biosynthesis, metabolism, and transportation. All 210 GT genes were identified from the pre-processed expression matrix, which can now be used to perform statistical analyses such as differential expression analysis to identify differentially expressed GT genes in each drug treatment case.

4.8 References

1. Ashburn, T. T., & Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. Nature reviews Drug discovery, 3(8), 673.
2. Carter, S. L., Brechbühler, C. M., Griffin, M., & Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics, 20(14), 2242-2250.
3. Cha, K., Kim, M.-S., Oh, K., Shin, H., & Yi, G.-S. (2014). Drug similarity search based on combined signatures in gene expression profiles. Healthcare informatics research, 20(1), 52-60.
4. Gillespie, C. S., Lei, G., Boys, R. J., Greenall, A., & Wilkinson, D. J. (2010). Analysing time course microarray data using Bioconductor: a case study using yeast2 Affymetrix arrays. BMC research notes, 3(1), 81.
5. Kaur, G., & Dufour, J. M. (2012). Cell lines: Valuable tools or useless artifacts: Taylor & Francis.
6. Kostka, D., & Spang, R. (2004). Finding disease specific alterations in the co-expression of genes. Bioinformatics, 20(suppl_1), i194-i199.
7. Lai, Y., Wu, B., Chen, L., & Zhao, H. (2004). A statistical method for identifying differential gene–gene co-expression patterns. Bioinformatics, 20(17), 3146-3155.
8. Lamb, J. (2007). The Connectivity Map: a new tool for biomedical research. Nature Reviews Cancer, 7(1), 54.
9. Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., . . . Ross, K. N. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 313(5795), 1929-1935.

10. Lee, M.-L. T., Kuo, F. C., Whitmore, G., & Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proceedings of the national academy of sciences, 97(18), 9834-9839.

11. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics, 28(6), 882-883.

12. Napolitano, F., Zhao, Y., Moreira, V. M., Tagliaferri, R., Kere, J., D'Amato, M., & Greco, D. (2013). Drug repositioning: a machine-learning approach through data integration. Journal of cheminformatics, 5(1), 30.

13. Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., & Tsui, K.-W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. Journal of computational biology, 8(1), 37-52.

14. Raychaudhuri, S., Stuart, J. M., & Altman, R. B. (1999). Principal components analysis to summarize microarray experiments: application to sporulation time series Biocomputing 2000 (pp. 455-466): World Scientific.

15. Sánchez, A., & de Villa, M. (2008). A tutorial review of microarray data analysis. Universitat de Barcelona.

16. Wang, G., Ye, Y., Yang, X., Liao, H., Zhao, C., & Liang, S. (2011). Expression-based in silico screening of candidate therapeutic compounds for lung adenocarcinoma. PloS one, 6(1), e14573.

17. Wang, K., Weng, Z., Sun, L., Sun, J., Zhou, S.-F., & He, L. (2015). Systematic drug safety evaluation based on public genomic expression (Connectivity Map) data: myocardial and infectious adverse reactions as application cases. Biochemical and biophysical research communications, 457(3), 249-255.

5 DIFFERENTIAL GENE EXPRESSION AND CO-EXPRESSION MODULE ANALYSIS

5.1 Microarray data analysis overview

The previous chapter described data preparatory steps for gene expression analysis. The raw expression data was collected and pre-processed. The resulting output is a gene expression matrix and the probe set IDs were annotated with gene names. The columns have drug treatment names and the rows have the gene names. Each drug treatment case had 5 drug treated samples and 5 untreated control samples. Where statistically required, some control samples were shared with amongst treated drug samples for consistent analysis. For instance, Alpha-estradiol and Alvespimycin shared common control sample. A standard R statistics package was used to build several design and contrast matrices and the LIMMA package from the Bioconductor was used to perform differential gene expression analysis.

A rank sum test was performed to investigate statistically significant drug treatment cases where the GT genes were differentially regulated compared with non-GT genes (Smyth, 2005) upon exposure to drug treatment. Visual data analytics such as Heatmaps were used to separate up from down regulated GT genes derived from the relative expressions for each drug treatment case. Following this, a co-expression module analysis was performed. The WGCNA package from the Bioconductor was used to identify significant modules of genes that are co-expressed from the selected drug treatment case from the rank sum meta-analysis. Finally, gene ontology analysis of the co-expressed modules was done to identify significantly enriched functional biological process of the co-expressed modules to draw conclusions. The analyses reported here were performed in an R (version 3.3.3) statistical programming environment via RStudio using the Bioconductor packages as detailed in chapter 2 section 2.3.3. An overall schematic view of microarray data analysis is given by figure 5.1.

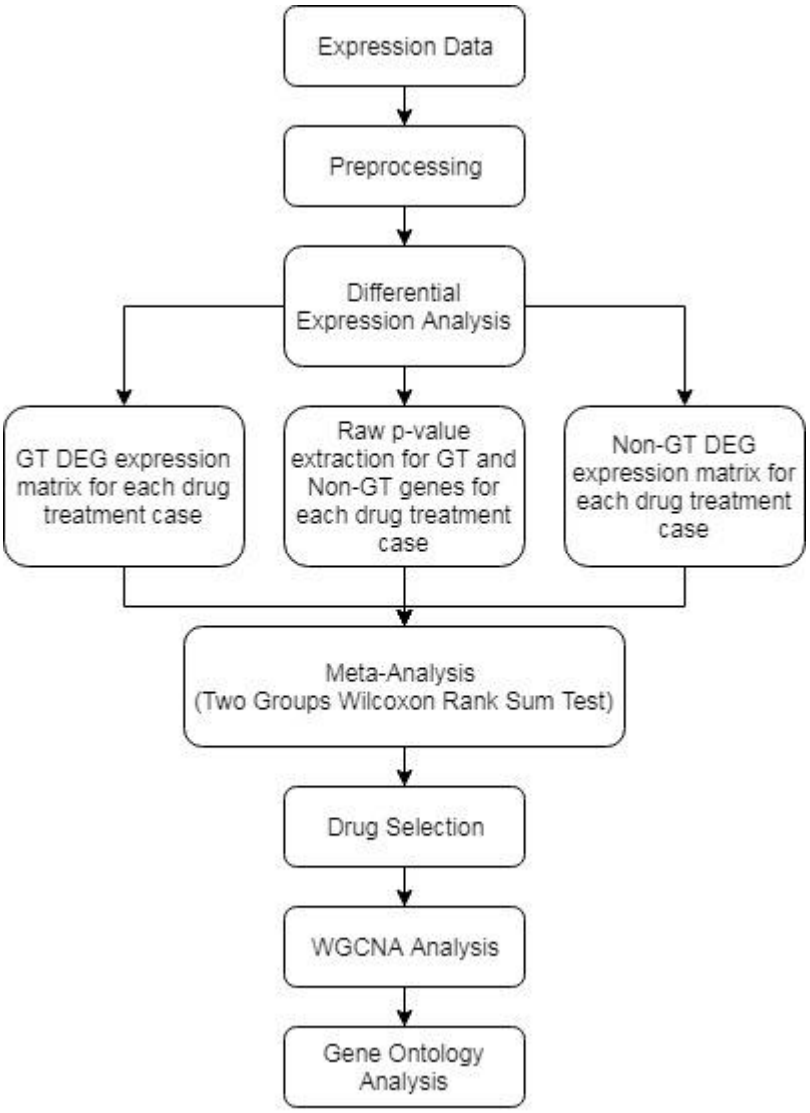


Figure 5.1 Schematic view of microarray data analysis process using CMap expression data.

5.2 Differential gene expression analysis

The main goal of a class comparison strategy from microarray data analysis is to identify and select the genes that have expression levels that are significantly different between contrasting experimental conditions. These genes are aptly defined as "differentially expressed genes". (Sánchez & de Villa, 2008) In the present study, genes that have a different expression level after drug treatment of a cell line compared to their expression in the case of no drug treatment are

defined as differentially expressed genes. Although this definition may appear trivial, there are several statistical measures that must be met for genes to be defined as differentially expressed. The methods and measures used to identify the differentially expressed genes for each drug-treated case from the pre-processed CMap gene expression matrix is detailed in this chapter within the framework of the LIMMA package. (Quackenbush, 2002)

5.2.1 LIMMA approach

LIMMA is a software package using linear models to analyse microarray gene expression data especially the analyses of designed experiments and the assessment of differential expression. LIMMA provides powerful means to simultaneously analyse contrasts between numerous RNA targets in designed experiments. LIMMA also offers an empirical Bayesian approach to produce stable results with a small number of sample replicates. (Ritchie et al., 2015)

The use of LIMMA was discussed in detail in chapter 2. There it was pointed out that a numerical gene expression matrix and one or two experimental design specific matrices are required for this analysis. First, a design matrix, that details the specific treatment sample relationships must be constructed. Second, a contrast matrix, that specifies which comparisons must be made between the expression samples is built. LIMMA then makes possible the fitting of gene-wise linear models which is used to estimate log-intensities between target RNA samples simultaneously to identify differentially expressed genes. Once the linear model is fitted using a constructed design matrix, a contrast matrix is used to compute log2 fold changes and finally t-statistics, allowing pairwise comparisons between the contrasts of samples of interest, are made. (Ritchie et al., 2015) (Smyth, 2005)

5.2.2 Design matrix

A design matrix was created using “model.matrix” function from the standard statistics package in R for all 29 drugs from the gene expression matrix. Both treatment and control samples were considered during the design matrix construction. The 203 representing total number of samples in the data as mentioned earlier, some drug-treatment samples shared a common control sample, that resulted in total number of treatment and control sample pair for 29 drugs of 42 (not 58). The

dimension of the design matrix was therefore 210x42 to which a linear model was fitted to the expression data.

5.2.3 Linear model fitting

Multiple linear models from generalized least squares were fitted using the "lmFit" function (LIMMA package). A linear model was fitted to a series of samples for each gene taken from each one of the 29 drug treatment cases. Following this linear fit coefficients, residuals, covariant coefficients and standard deviations were used to describe the differences between the RNA samples hybridized to the arrays. (Ritchie et al., 2015)

5.2.4 Contrast matrix

After a linear model was fitted for each gene across the samples the "makeContrasts" function was used to construct a contrast matrix that generates contrasts between a set of parameters in a numeric matrix. The parameters are represented by the coefficients from linear model fit. The contrast matrix reveals the comparisons that are made between the coefficients and the coefficients that are extracted from the linear model fit. (Ritchie et al., 2015) (Smyth, Ritchie, Thorne, & Wettenhall, 2005) In this study, the contrast matrix was constructed for drug-treated samples and control samples. For example, in the Alpha-estradiol treatment, the contrast names were defined as "Alpha-estradiol" and "MCF-7 control". This was the case for all 29 selected drugs. The contrasts from the linear model fit was computed using the "contrasts.fit" function and the fitted model object, from the coefficient of the design matrix, was re-oriented to the corresponding set of contrasts of the original coefficients following Smith et al.'s protocol. (Smyth et al., 2005)

5.2.5 Empirical Bayes fit

To statistically assess the differential gene expression, LIMMA uses a parametric empirical Bayes method (described in chapter 2) to moderate the standard errors of the estimated log-fold changes that results in i) more stable and robust inference and ii) improved power in analysis for a small number of samples. (Smyth, 2004) (Phipson, Lee, Majewski, Alexander, & Smyth, 2016) Here the function "eBayes" was used to compute moderated t-statistics, moderated F-statistics, and log-

odds of differential expression. Various coefficients were returned as output values. The empirical Bayes adjusted expression matrix with contrast design was then available to extract the top most differentially expressed genes.

5.2.6 Extraction of top differentially expressed genes

The "topTable" function was applied to identify differentially expressed genes for each linear model fitted to the 29 drug treatment contrast case. A table of top-ranked differentially expressed genes from the linear model fit and various design and contrast matrices were then extracted by tuning the "topTable" function settings which we describe here.

The function returns $\log_2$ FC (fold change), average expression values, moderated t-statistics, p-value, adjusted p-value and B statistics in its output. The $\log_2$ FC value is an estimate of the $\log_2$-fold-change that corresponds to the contrast value (see section 5.2.4). The average expression returns the i) average $\log_2$-expression for the probe over all sample arrays, ii) the raw p-values, iii) the false discovery rate (FDR) adjusted p-values and iv) the B statistics which is a log-odds that the gene is differentially expressed. Controlling the FDR as described in chapter 2 is essential for multiple hypothesis testing analyses such as differential gene expression analysis. The p-values for the contrast of interest were adjusted for multiple testing (using "p.adjust" in R) and the Benjamini & Hochberg method was used to control the expected FDR below the specified value.

The threshold of adjusted p-value < 0.05 and |log2 FC (fold change) | > 0 and $\log_2$ FC < 0 were set to maximize the output of the differentially expressed genes in so doing improving the precision of gene selection. The genes that fall into the log2 FC range from -0.5 to 0.5 show no difference in expression. However, when the expression meets the p-value cut off < 0.05, those genes are identified as differentially expressed. All genes whose expression levels met these requirements were selected for subsequent analysis. For each drug treatment case, the genes were divided into two modules. The differential genes with $\log_2$ FC < 0 were labelled down-regulated gene module and the differential genes with $\log_2$ FC > 0 were labelled up-regulated gene module.

Overviewing, the entirety of the statistical analysis performed in this thesis it will become apparent that the analyse can be grouped into one of two phases. The aim of the first phase was to identify differentially expressed GT and non-GT genes from the drug treatments. The aim of the second

phase was to identify the drug treatment that significantly influenced the regulation of GT genes in comparison to non-GT genes using the results from differential expression analysis. Following the second phase a set of selected drugs is used to construct a gene co-expression network with gene modules to make an inference about biological functions associated with specific gene co-expression modules.

5.3 Identifying differentially expressed genes

Following the protocol outlined above a set of differentially expressed genes was returned for all 29 drug treatments. The specific choices for FDR control with a threshold of adjusted p-value < 0.05, $\log_2$ FC > 0 and $\log_2$ FC < 0, 26 for the drug treatment cases gave statistically validated differentially expressed genes. Of the 26 drug treatment case with DEGs a subset of 19 drug treatment cases, produced DEG that included differentially expressed glycosyltransferase (GT) genes. The information listing the DEGs and differentially expressed GT genes for each drug treatment case is listed Table 5.1.

Several cancer drugs, including breast cancer drugs were used in the CMap project. It was predicted that the anti-cancer drugs would generate the most DEGs because these drugs would suppress the cancerous genes also, these drugs were tested on MCF-7 breast cancer cell line and indeed, the most DEGs and differentially expressed GT genes were produced by the anti-cancer drug treatment. I now consider the three drugs that produced the largest DEGs and GT-DEGs. The largest number of differentially expressed genes (6066 DEGs) were retrieved following the Vorinostat drug treatment. This drug was originally developed as a histone deacetylase inhibitor (HDAC) for the treatment of cutaneous T-Cell Lymphoma (CTCL). In contrast the lowest DEGs resulted from pioglitazone and rosiglitazone treatment that are type II diabetes mellitus treatment drugs. While Vorinostat was primarily designed for the treatment of CTCL its usage has been extended to the treatment of breast cancer. (Munster et al., 2011) (Luu et al., 2008). Second on the list is Trichostatin (TSA) which is an organic compound with that has been considered as a possible anti-cancer drug. TSA promotes over-expression of genes that initiate apoptosis in a cancer cell to slow down the cancer progression. (Drummond et al., 2005) Then in the list of high DEGs follows a group of chemically related drugs (Tanespimycin, Geldanamycin and Alvespimycin) that have been used as dual purpose antibiotics and cancer treatments. Unsurprisingly they yield the same

levels of differential expression. Tanespimycin is a derivative of the antibiotic geldanamycin that is used to treat young patients with leukaemia and other tumours. (Dimopoulos, Mitsiades, Anderson, & Richardson, 2011) Alvespimycin is a derivative of Geldanamycin and inhibitor of heat shock protein (HSP) 90 together with Monorden, that is used to treat solid tumours in various type of cancers. (Pacey et al., 2011) Tretinoin is an all-trans retinoic acid (ATRA) that is effective for the treatment of acute promyelocytic leukaemia. (Warrell Jr et al., 1991)

Table 5.1 Differentially expressed GT and non-GT genes identified for each drug treatment.

Drug Name	DEGs identified	GT genes identified
Vorinostat	6066	75
Trichostatin A	4920	60
Geldanamycin	3339	37
Alvespimycin	3155	37
Tanespimycin	2695	27
LY-294002	2540	31
15-Delta Prostaglandin J2	2416	32
Thioridazine	2156	17
Fulvestrant	1903	16
Monorden	1749	19
Sirolimus	1711	17
Trifluoperazine	1434	16
Monastrol	1271	15
Tretinoin	1236	13
Wortmannin	1152	11
Prochlorperazine	1147	12
Fluphenazine	692	4
Genistein	157	3
Estradiol	49	1
Valproic Acid	15	0

Alpha-estradiol	19	0
Troglitazone	9	0
Rosiglitazone	0	0
Pioglitazone	1	0
Nordihydroguaiaretic Acid	0	0
Clozapine	0	0
Chlorpromazine	2	0
Acetylsalicylic Acid (Aspirin)	8	0
Haloperidol	13	0

Finally closing out the list of high DEGs is LY-294002, an inhibitor of phosphoinositide 3-kinases (PI3Ks), a group of enzymes that are involved in the regulation of numerous cellular functions. LY-294002 has been identified as an effective anti-cancer agent (Maira, Stauffer, Schnell, & García-Echeverría, 2009). 15-Delta Prostaglandin J2 is a subset of prostaglandins (PGs) that exerts diverse biological functions, such as suppression of inflammation responses, growth, and survival of cells. 15-Delta Prostaglandin J2 has been identified and modelled for the development of novel anti-cancer drugs. (Wang & DuBois, 2006) (Clay, Monjazeb, Thorburn, Chilton, & High, 2002)

Interestingly much further down the list (Table 5.1). Fulvestrant is not in the group of high DEGs and GT-DEGs although it was designed to treat hormone receptor (HR)-positive metastatic breast cancer in postmenopausal women with disease progression and is widely used in clinical treatment. (Lee, Goodwin, & Wilcken, 2017) Another breast cancer drug that is not in the high DEGs set is Wortmannin. This is a steroid metabolite that is an inhibitor of PI3Ks and is known to inhibit a proliferation of MCF-7 breast cancer cell lines. (Karve et al., 2012) (Yun et al., 2012) Bottoming out the list is another cancer drug Genistein. This drug is an isoflavone that is both an antioxidant and anthelmintic, that has been tested and used as an effective in the treatment of prostate cancer. (Banerjee, Li, Wang, & Sarkar, 2008)

Table 5.2 An output of differential gene expression analysis from Vorinostat treatment displaying top 5 up and down regulated GT genes.

	Gene	log_2 FC	Ave Expr	t-statistics	p-value	Adjusted p-value
Up-regulated	NEU1	1.7985	8.4820	22.6399	$9.1086e^{-54}$	$1.1328e^{-51}$
	EXTL2	1.4230	6.7741	16.8145	$2.5692e-^{38}$	$9.2084e^{-37}$
	FUCA1	1.2590	8.3014	18.9469	$3.5853e^{-44}$	$2.0360e^{-42}$
	C1GALT1	1.0925	6.3813	14.2090	$6.3511e^{-31}$	$1.2970e^{-29}$
	B4GALT5	0.9175	8.3839	13.5088	$6.5615e^{-29}$	$1.1809e^{-27}$
Down-regulated	ST6GALNAC	-1.3139	7.8520	-16.0120	$4.6106e-^{36}$	$1.3884e^{-34}$
	B3GALNT1	-1.2487	8.4222	-13.1294	$8.1344e^{-28}$	$1.3435e^{-26}$
	PIGB	-0.9447	7.1891	-11.6572	$1.3964e^{-23}$	$1.6780e^{-22}$
	ST3GAL1	-0.7697	6.3247	-6.8084	$1.5340e^{-10}$	$7.4965e^{-10}$
	MAN2A2	-0.7671	6.2627	-13.2427	$3.8354e^{-28}$	$6.4987e^{-27}$

The drugs discussed above are primarily used to treat various type of cancers. Besides cancer drugs, some non-cancer drugs such as sirolimus and prochlorperazine interestingly produced a significant number of DEGs and differentially expressed GT genes. Given that Vorinostat tops the list with the most DEGs and DEG-GTs it appears as having the greatest potential to regulate the GT genes involved in breast cancer progression. It is worthwhile to consider a typical output of differential gene expression analysis (Table 5.2).

5.4 Using meta-analysis to select effective breast cancer drugs

While a simple numerical ranking of number of DEGs is appealing a more thorough approach is needed to select the drugs that have the most significant effect on changing the GT gene regulation from that of the untreated breast cancer cell line. To achieve this a meta-analysis with two-groups

Wilcoxon rank sum test was performed using the raw p-values and gene expression matrices of the selected drug treatment obtained after differential expression analysis. Statistical methods we therefore used to investigate and select drugs that had greater influence on the regulation of GT genes (either up or down regulation) than on the regulation of non-GT genes. A subsequent visualization, co-expression module and gene ontology analyses were done based on the selected drugs.

Meta-analysis is a statistical approach that systematically evaluates the pooled data from separate studies to measure for statistical significance that will point to unbiased. Typically, performing a differential gene expression analysis with a limited number of biological replicates would potentially increase the generation of false positives and this would lead to a misinterpretation of statistical significance. Including a meta-analysis step in the statistical workflow improves the robustness of the process and delivers greater reliability. (Allemeersch & Moreau, 2004) In this study it provides a greater confidence in the selection of drugs that are important in affecting GT gene regulation compared with the regulation of non-GT genes.

A robust rank sum test in the form of meta-analysis to select drugs that had significant influence in the regulation of GT genes will now described. The test was performed using all the outputs of differential expression analysis of the 29 drug treatments and so overcoming the potential problem of a limited sample size in microarray data analysis.

5.4.1 Two-groups Wilcoxon rank sum test

A typical differential gene expression analysis would return several matrices containing information about the $\log_2$ FC (fold change), t-statistics, p-values and adjusted p-values for example as described in section 5.2.6 and the p-values returned is very useful for rank sum analysis. As previously stated the goal was to measure if GT genes are more significantly regulated in comparison to non-GT genes. Consequently, a ranking analysis of the GT and non-GT genes would be suitable for the comparison method. For this reason, a traditional non-parametric Wilcoxon rank sum test (also known as, Mann-Whitney U test) was used. (Troyanskaya, Garber, Brown, Botstein, & Altman, 2002) In this case only a two-group rank sum test was necessary since a rank comparison between the GT and non-GT genes were being made. The gene expression matrix after pre-processing had dimension of 22,215 by 203 as described in Chapter 4. However,

after removal of the duplicated probe ID set representing a single gene name, the dimension of gene expression matrix was reduced to 12,436 by 203. All the subsequent meta-analysis used this reduced expression matrix.

5.4.2 Methodology

Initially, the gene expression matrix described (see section 5.3.1) was used to run a differential gene expression analysis. The analysis returned p-values as well as adjusted p-values in the output table. Since the aim was to rank the genes based on the p-values, a set of non-adjusted p-values were used to construct a p-values matrix for each drug treatment experiment. In total, 29 drug treatment p-values matrices were constructed, and they were merged into a single p-value matrix representing all 29 samples. Since there are 12,436 p-values, representing each gene and 29 drug treatment case (each drug treatment had 10 samples including treatment and control and this had gone through differential gene expression analysis) the dimension of the p-value matrix was 12,436 by 29.

Next, for each drug treatment, the p-values matrix was divided into a p-value matrix for GT genes and p-value matrix for non-GT genes through data manipulation techniques. Since the number of GT genes were a lot less than the non-GT genes, random sampling of non-GT genes without replacement was used to select the number of non-GT genes that matched the number of GT genes. For instance, if there were 75 GT genes from the p-values matrix A, the same number of non-GT genes were selected using random sampling without replacement method out of all the non-GT genes. The randomly selected non-GT genes with its respective p-values together with the selected GT genes with its respective p-values were used to perform two-group Wilcoxon rank sum test that works out the rank of individual p-values and evaluates if the GT genes were significantly regulated in comparison to the non-GT genes by that specific drug treatment. The two-groups Wilcoxon rank sum test that required random sampling of non-GT genes was done 1,000 times for each drug treatment to increase the statistical significance. An R function was created to perform this sampling procedure.

To better understand the process, a simple R function that performs random sampling and two-groups Wilcoxon rank sum test between GT and non-GT genes is shown below.

```
wilcox_function <- function(gt_matrix, non_gt_matrix) {
    # apply 1 to 29 for each sample in columns
    gt <- gt_matrix[, 1]
    # define an empty vector
    m <- c()
    # iterate over1000 times
    for (i in 1:1000) {
        # apply 1 to 29 for each sample in columns
        non_gts <- non_gt_matrix[sample(1:nrow(non_gt_matrix), 143, replace = FALSE), 1]
        rank_sum_result <- wilcox.test(gt, non_gts)
        # extract p-values
        rank_p_value <-  rank_sum_result$p.value
        m <- c(m, rank_p_value)
    }
    # return a vector of 1,000 p-values from Wilcoxon rank sum test
    return(m)
}
```

Typically, the two-groups Wilcoxon rank sum test returns p-value for the rank sum test and if the p-value of the test returned is less than a decided threshold, for instance, p-value < 0.05, the test is deemed significant, and the null hypothesis of "*GT genes are not significantly regulated by the drug*" will be rejected. This implies that the drug treatment has had a significant effect on the regulation of the GT genes than on the non-GT genes. This protocol made possible a selection of drug treatments that had greatest effect in the GT genes. These were then used for co-expression module and gene ontology analysis to connect the treatment result with a molecule mechanism. The meta-analysis workflow is summarized in figure 5.2.

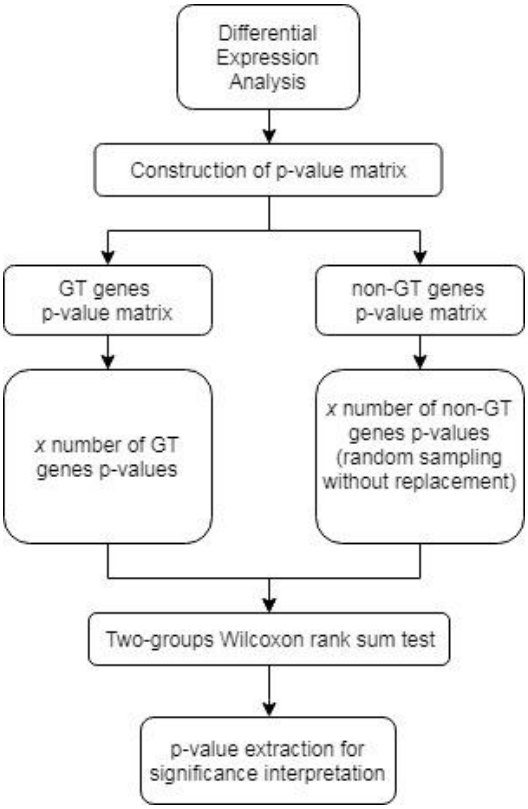


Figure 5.2 Schematic view of meta-analysis deploying two-groups Wilcoxon rank sum test. For each drug, x number of non-GT genes p-values were selected by random sampling 1,000 times. Two-groups Wilcoxon rank sum test was performed 1,000 times between selected GT and non-GT p-values.

5.4.3 Drug selection

The drugs that had greatest effect in the regulation of GT genes were selected based on the 1,000 p-values reported in the meta-analysis (section 5.3.2.). After a set of p-values were returned from meta-analysis of each drug, a histogram of the p-values was plotted for each drug. The distribution of the p-values was observed and the frequency of the p-values at different significance thresholds were observed. Threshold p-value < 0.05 was selected statistically significance. Thus, the more p-values that were observed from the range 0 to 0.05 in the histogram, the more statistically significant the rank sum test was. Additionally, if the distribution of the histogram was skewed to the right, meaning each round of rank sum test produced lower p-values, the drug was considered to have had a significant effect on regulating the GT genes in comparison to the non-GT genes.

The drugs were selected based on two criteria. Firstly, the histogram's distribution should have a skewness to the right and the frequency of p-values must be observed in the range 0 to 0.05 or at least those in the 0.1 region should be high. If this criterion is satisfied as a result of the two-groups rank sum test that was performed for 1,000 times for a drug treatment using randomly sampled non-GT genes, and the p-values observed were more in the significance threshold region, then it likely that the drug had a significant effect on the regulation of the GT genes. A second criteria is that the drug must have produced any differentially expressed genes (DEGs) as part of its output. If the drug did not produce any DEGs in the first place but showed significant regulation of GT genes from meta-analysis, this drug was not considered for subsequent co-expression and gene ontology module analysis. The two-groups Wilcoxon rank sum test results in the form of histogram for each of selected cancer and non-cancer drugs are available in figure 5.3.

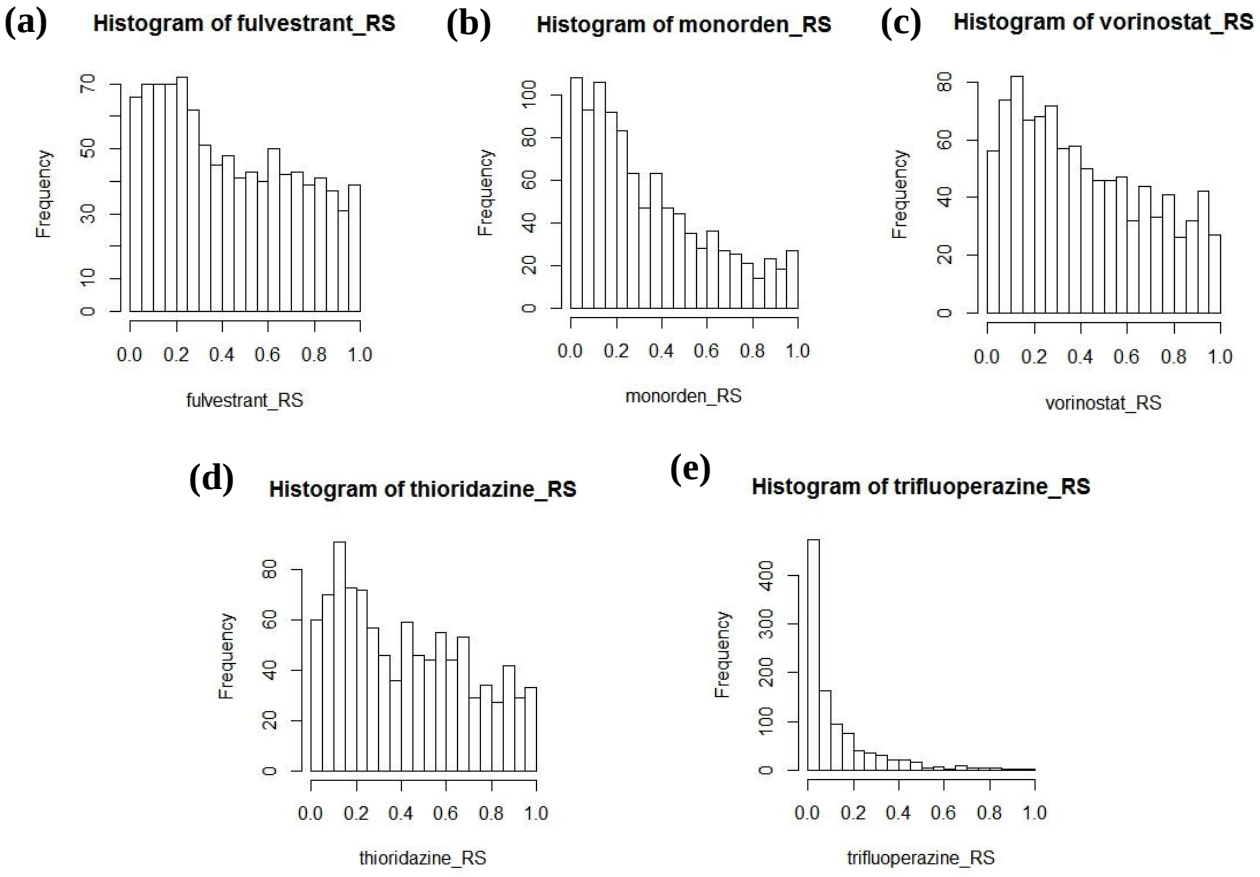


Figure 5.3 Distribution of p-values obtained from the meta-analysis for cancer and non-cancer drugs. A significance threshold of p-value < 0.05 was selected for the Rank Sum Test. The drug selection was based on the distribution (in terms of skewedness) of the histogram and the frequency of p-value observed in the 0 to 0.05 region. 5 drugs including, (a) Fulvestrant, (b) Monorden, (c) Vorinostat, (d)

thioridazine and (e) trifluoperazine were selected for further analysis. The remaining histograms for another drug are available in the graph appendix B.

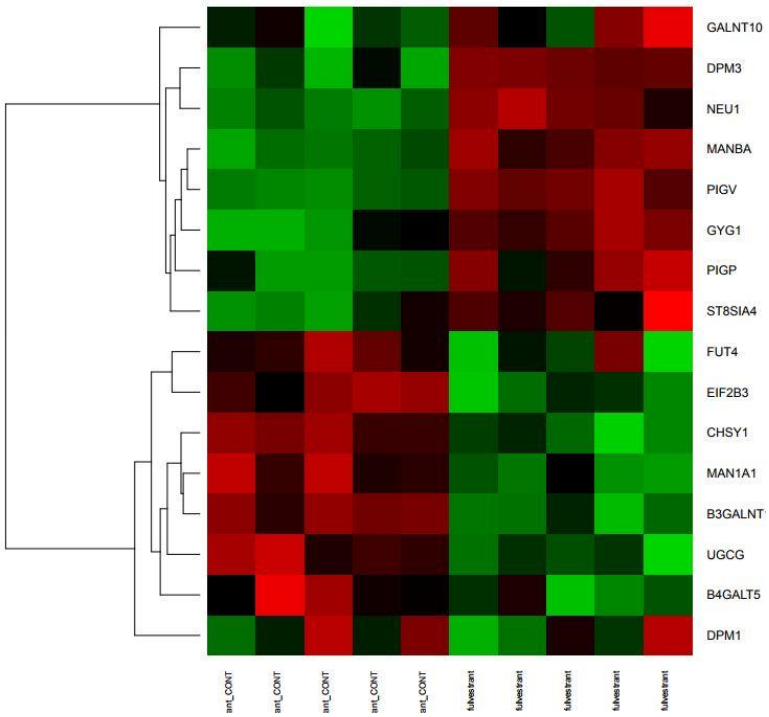
The 29 different histograms were plotted for each drug treatment and through an inspection of p-value distribution and frequencies 8 candidate drugs were selected to have a significant effect in the regulation of the GT genes, namely, Fulvestrant, Monorden, Nordihydroguaiaretic acid, Rosiglitazone, Thioridazine, Trifluoperazine, Troglitazone and Vorinostat. Of the 8 candidate drugs, only 5 were found to have produced any DEGs (see Table 5.1). These drugs were Fulvestrant, Monorden, Thioridazine, Trifluoperazine and Vorinostat. They were selected for subsequent analysis such as the co-expression module and gene ontology. Next the 5 drugs are considered in one of two groups either a cancer drug or non-cancer drug.

5.4.3.1 Cancer drugs

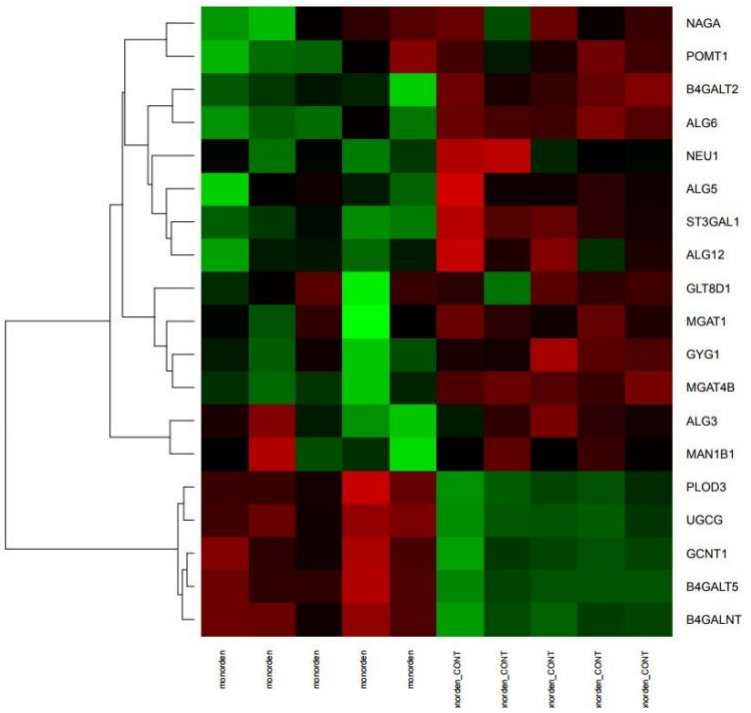
Remarkably 3 of the 5 drugs were existing cancer drugs that arose from their being selected based on their statistically significant performance following MCF-7 treatment that led to substantive altered regulation of GT genes compared with non-GT genes. The cancer drugs are Fulvestrant, Monorden, and Vorinostat. Fulvestrant is a medication used to treat hormone receptor (HR)-positive metastatic breast cancer in postmenopausal women. (Nathan & Schmid, 2017) Monorden, also known as Radicicol is a substance that binds to heat shock protein 90 (HSP 90), which is a chaperon protein that stabilizes proteins involved in cancer tumour growth and altering its function. HSP 90 inhibitors are frequently studied and investigated as anti-cancer drugs because of its pharmacological inhibition ability in HSP 90. HSP 90 inhibitors have great potential as anti-breast cancer drug as they can suppress multiple oncogenic signalling pathways simultaneously. This effectively reduces possible molecular feedback loops and tumour resistance mutations. (Zagouri et al., 2013) As was mentioned previously Vorinostat is an immune system cancer drug that is a HDAC inhibitor used for treatment of CTCL. Of greater relevance here, Vorinostat has been well-studied and investigated for breast cancer treatment because of its noted potential as an anti-breast cancer drug. (Munster et al., 2011) (Luu et al., 2008)

A heatmap is an effective visualization technique that is frequently used in genomic studies to visualize gene expression data. Here heatmaps displaying differentially expressed GT genes before and after treatment is shown for Fulvestrant, Monorden and Vorinostat (Figure 5.4).

(a)



(b)



(c)

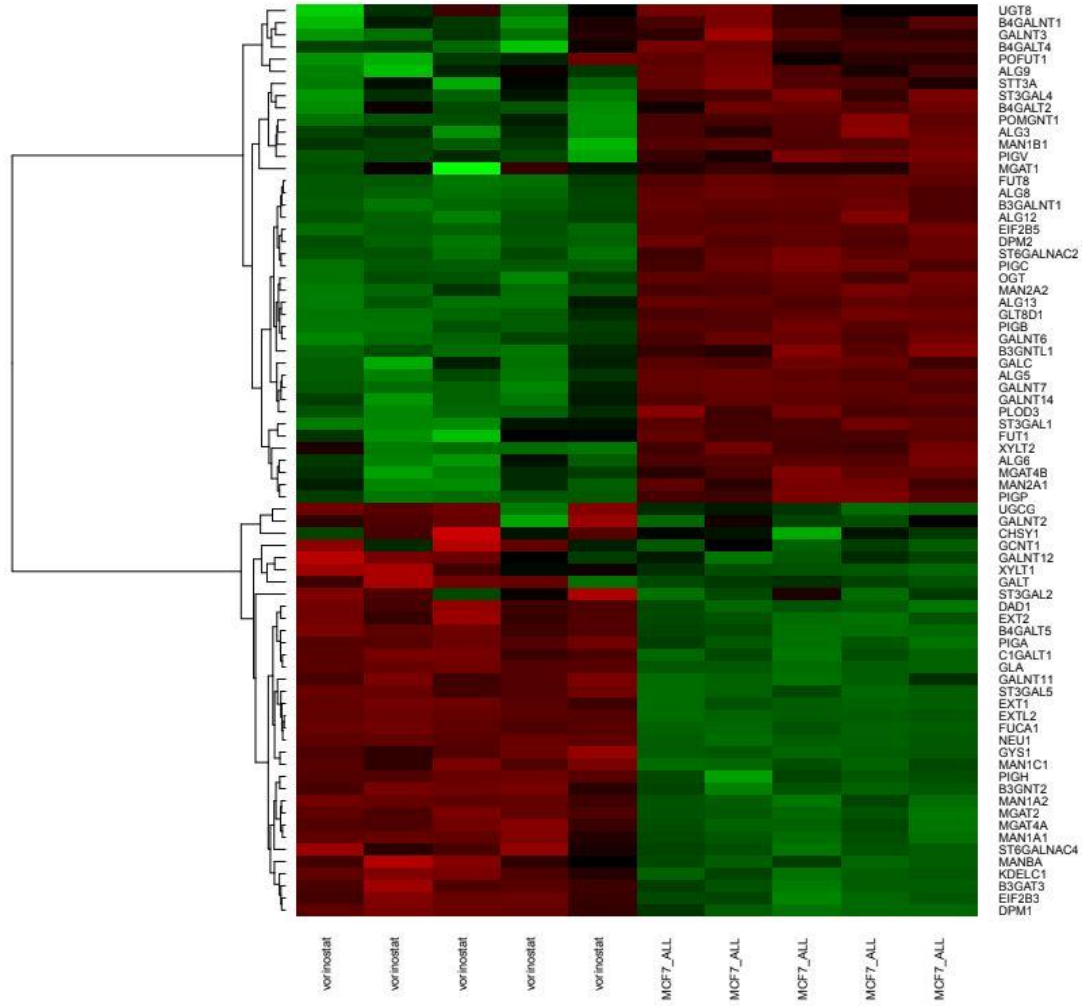


Figure 5.4 Heatmaps showing glycosyltransferase genes expression pattern before and after treatment of anti-cancer drugs on the MCF-7 breast cancer cell line. Each drug had (n = 10) samples in which 5 were treatment and the other 5 were control samples. Clear separation of up and down regulation of the genes can be identified from the heatmap visualization. (a) Fulvestrant treatment: left 5 are control and right 5 are treatment. (b) Monorden treatment: left 5 are treatment and right 5 are control. (c) Vorinostat treatment: left 5 are treatment and right 5 are control. Expression profiles were clustered using hierarchical clustering with agglomeration method of "ward.D2".

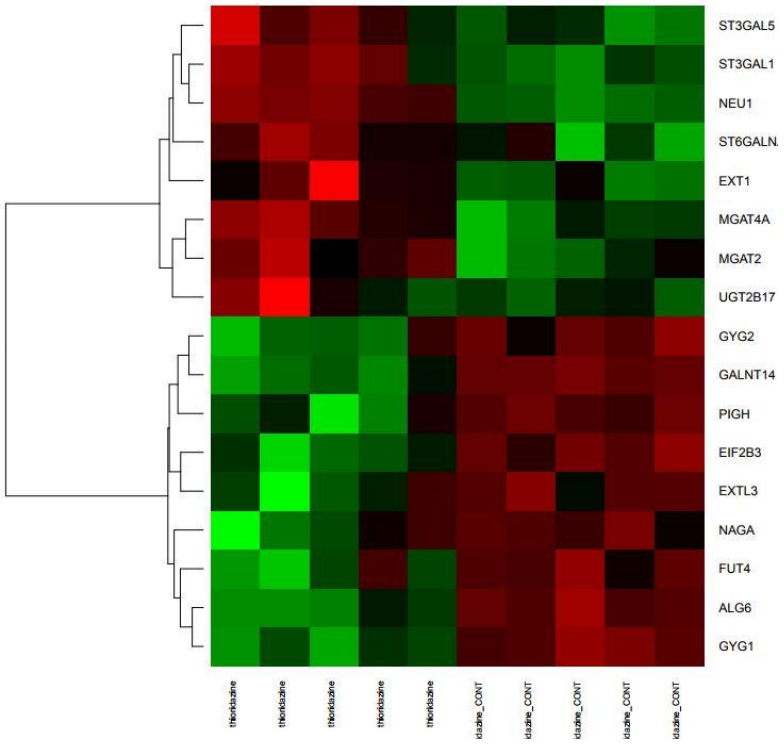
The heatmaps of anti-cancer drugs show various differentially expressed GT genes regulation pattern. For instance, following Fulvestrant treatment, some important GT genes such as FUT4 and B4GALT5 have been down-regulated after Fulvestrant treatment whereas, they were up-regulated in MCF-7 cells without treatment. This was anticipated as Fulvestrant is a metastatic breast cancer treatment drug. Monorden treatment resulted in the down-regulation of important

GT genes involved in breast cancer such as ST3GAL1 (see Figure 5.4 (b)). Studies have revealed that over-expression of ST3GAL1 promotes mammary tumorigenesis. (Picco et al., 2010) However, B4GALT5 which was down-regulated with Fulvestrant treatment was up-regulated with Monorden treatment. The Vorinostat treatment which generated the most differentially expressed GT genes (n = 75) results in the down-regulation of a number of different families of GT genes involved in breast cancer, such as FUT families and ST3GAL families. ST3GAL1 gene was up-regulated in the MCF-7 cell line without treatment however, after Vorinostat treatment, it was down-regulated (Figure 5.4 (c)).

5.4.3.2 Non-cancer drugs

There were 2 non-cancer drugs namely, thioridazine and trifluoperazine. Both thioridazine and trifluoperazine are antipsychotic drugs primarily used to treat schizophrenia. (Hedberg, Houck, & GLUECK JR, 1971) (Overall, Hollister, Meyer, Kimbell, & Shelton, 1964) Some studies have used thioridazine and trifluoperazine to inhibit breast cancer growth. For instance, Wei et. al. (Wei, Hickie, & Klaassen, 1983) studied the effect of trifluoperazine on the growth of breast cancer cells and revealed in their in vitro study that trifluoperazine had the ability to inhibit the colony formation from the breast cancer cell line. (Strobl & Peterson, 1992) Further it was shown that thioridazine could utilize a distinct breast cancer pathway to inhibit the cell growth, leading to the conclusion that this is a potential noncytotoxic alternative to tamoxifen for treatment of tamoxifen-resistant breast cancer. The heatmaps of non-cancer drug displayed in Figure 5.5.

(a)



(b)

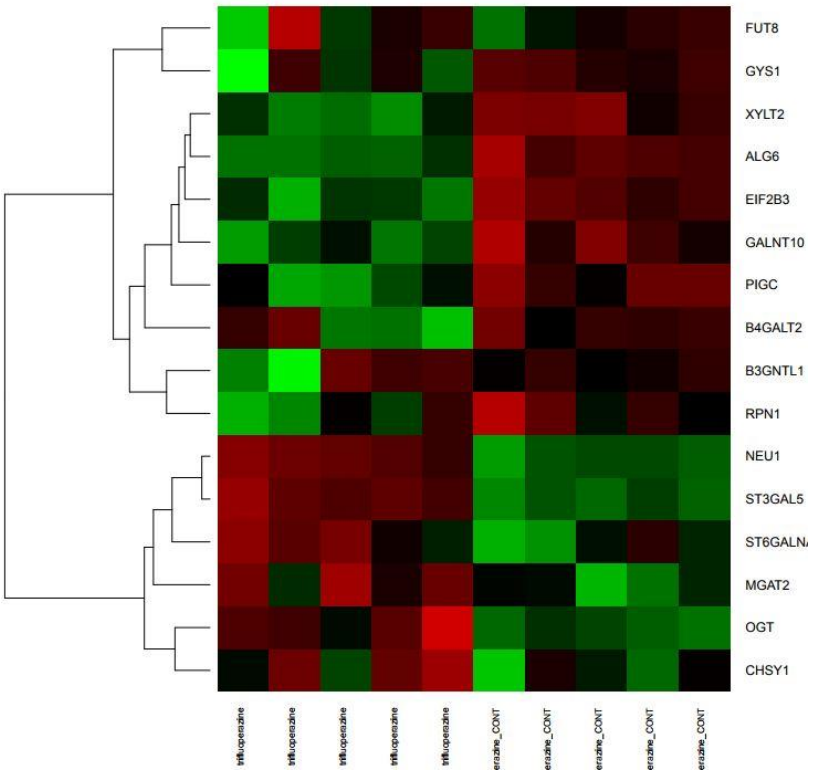


Figure 5.5 Heatmaps showing glycosyltransferase genes expression pattern before and after treatment of non-cancer drugs on the MCF-7 breast cancer cell line. Each drug had (n = 10) samples

in which 5 were treatment and the other 5 were control samples. Clear separation of up and down regulation of the genes can be identified from the heatmap visualization. (a) Thioridazine treatment: left 5 are treatment and right 5 are control. (b) Trifluoperazine treatment: left 5 are treatment and right 5 are control. Expression profiles were clustered using hierarchical clustering with agglomeration method of "ward.D2".

Thioridazine drug unfortunately does not show a down-regulation of some of the ST3GAL family of genes including ST3GAL5 and ST3GAL1 and as well as members of the MGAT families including MGAT4A and MGAT2 genes. These genes were up-regulated after treatment of thioridazine. Surprisingly the FUT4 gene is down-regulated after thioridazine treatment, though in one sample, it has been up-regulated. This probably would indicate that thioridazine may not be an effective treatment for breast cancer since it did not down-regulated some of the important GT genes involved in breast cancer proliferation. Trifluoperazine's effectiveness in breast cancer suppression is also questionable judging from the heatmap in figure 5.5 (b). GT genes such as NEU1, ST3GAL5, MGAT5 have all been up-regulated after the treatment. Although a few genes such as FUT8 and B4GALT2 have shown down-regulation pattern after treatment, in some samples, they have been up-regulated and this could indicate that trifluoperazine may not be the most effective treatment for breast cancer. It is expected that if the treatment was effective, across all 5 samples, the tumour cells should have same expression patterns as normal breast cells, but this was not the case in trifluoperazine treatment.

5.5 Co-expression analysis

Next a search for clues to GT gene biological function is undertaken through co-expression gene module analysis using the Weighted Gene Co-Expression Network Analysis (WGCNA) method. The co-expressed modules of genes are identified, and they are associated with specific biological processes through a gene ontology analysis. Attention is paid to GT genes that were up and down-regulated after specific drug treatment and identified in a co-expressed module in so doing revealing the biological process with which that module associated with. From this analysis it is hoped that an insight into how the GT genes that are co-express with other genes including oncogenes to regulate a specific biological function in the development of tumour cells. Gene co-expression network analysis aim to identify modules of genes that show a coordinated expression pattern based on a pairwise correlations between genes. (van Dam, Võsa, van der Graaf, Franke,

& de Magalhães, 2017) The WGCNA, a widely used systems biology method was used to construct a scale-free network from gene expression data of selected drug treatment from section 5.3.3. The next few paragraphs describe how WGCNA was performed on the selected drug treatment to identify co-expressed modules of genes for subsequent gene ontology analysis.

5.5.1 Weighted gene co-expression network analysis approach

The WGCNA was performed using all the differentially expressed genes (DEGs) from section 5.2.7 for drugs selected in section 5.3.3. Performing WGCNA using the glycosyltransferase gene subset would not be appropriate approach since the sample size is too small compared to all the DEGs generated. The analysis was performed in R using the WGCNA package which comprises functions that perform a correlation network analysis to cluster similarly expressed genes into a module.

Following the WGCNA best practices (Langfelder & Horvath, 2014), for each gene expression data from selected drugs, the genes with zero counts or low variance were removed using the goodSampleGenes function. The co-expression analysis was initiated by calculating Pearson's correlation matrices for all pairs of genes. The correlation coefficient between gene m and gene n was defined as Smn = |cor(m,n)|. The soft thresholding power (β) was computed (pickSoftThreshold function). The model fitting of expression dataset to a scale free topology is shown below (Figure 5.6).

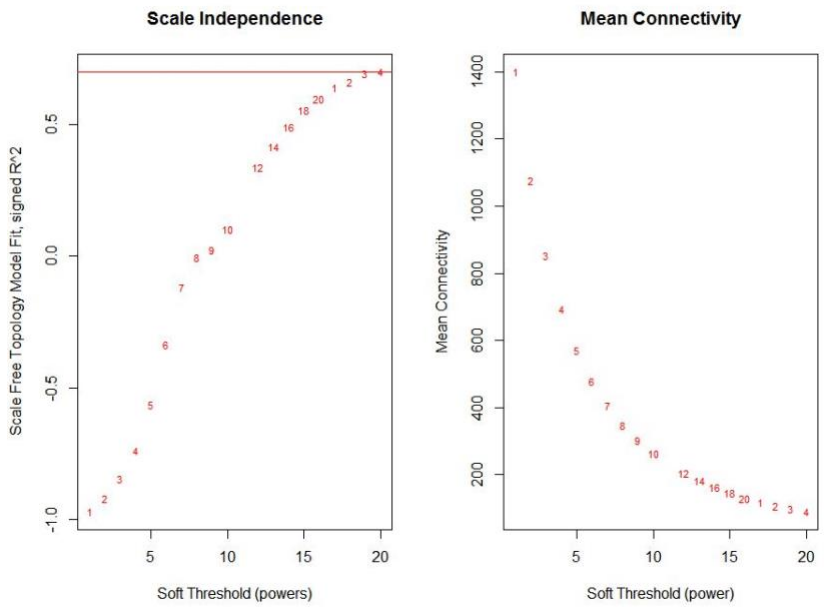
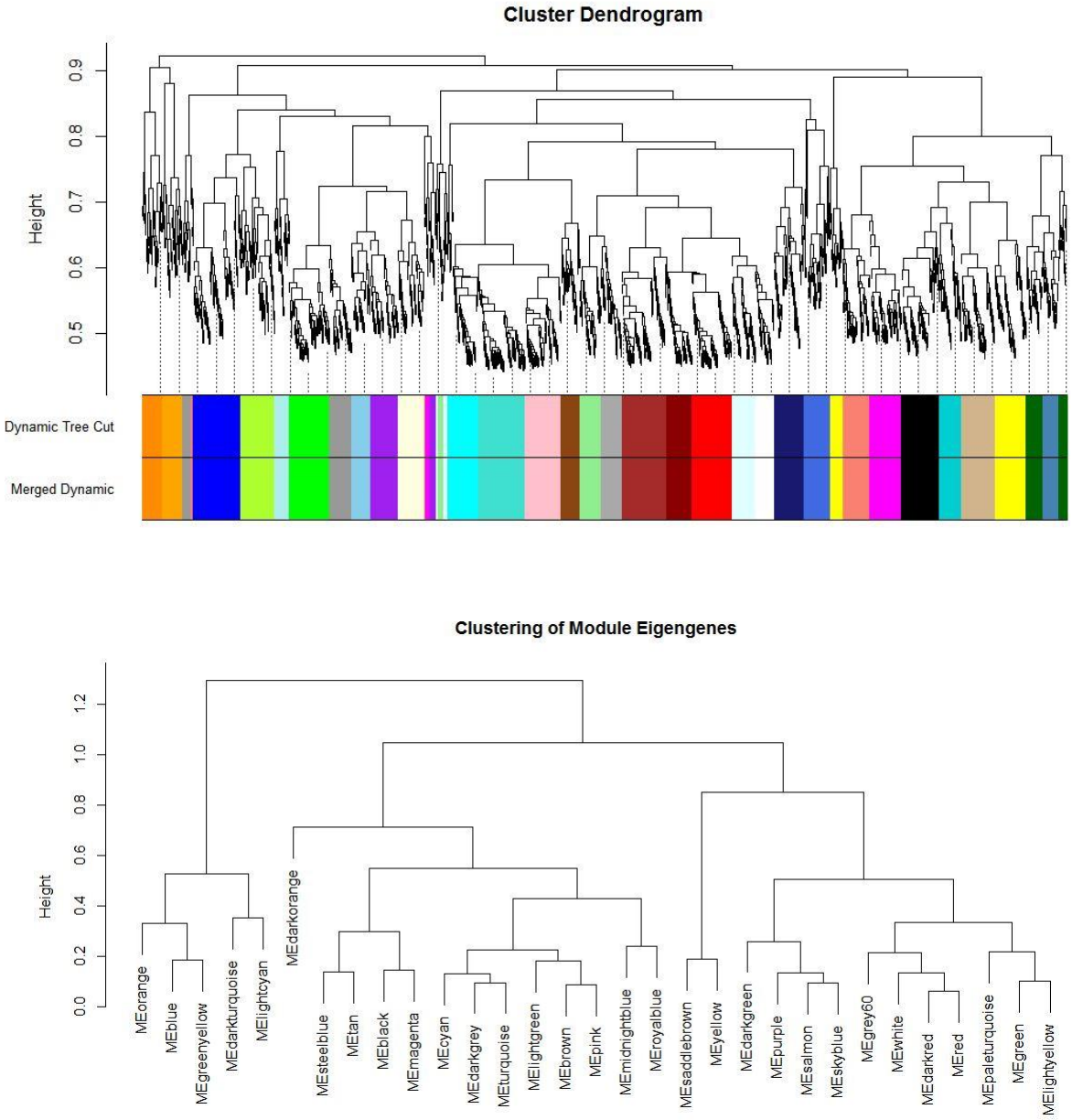


Figure 5.6 Network fitting to the scale-free topology accordingly to the parameter used in the soft-thresholding. The linear model of the regression line, Soft.R.sq was 0.65 with chosen β = 4.

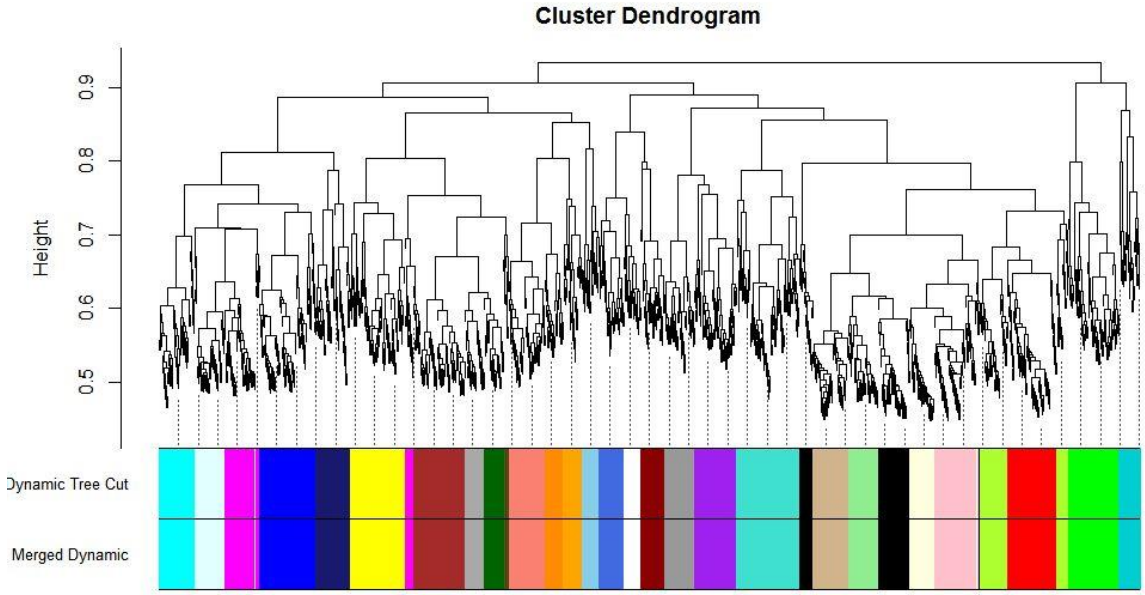
According to figure 5.6, that represents soft thresholding linear model of the regression line, the power of 4 was chosen since it was the lowest power for which the scale free topology index reached 0.65 (red line). However, the author of WGCNA as a caveat mentions that sometimes scale free topology cannot be reached for reasonably low values of β due to severe array outliers or globally distinct groups of arrays that lead to strong correlation between the expression profiles (and very large co-expression modules). In this case, for an unsigned network, β of 6 used to obtain a scale-free network. (Langfelder & Horvath, 2008) (Langfelder & Horvath, 2014) Next, the Pearson's correlation matrices were transformed into adjacency matrices by computing absolute value of pairwise Pearson's correlation between gene expression profiles each raised to the soft threshold power of β = 6. The power function a_{mn} = power (S_{mn}, β) = $|S_{mn}|^{\beta}$. The adjacency matrices were transformed topological overlap measure (TOM) similarity matrices that measures a gene pair connectivity. A measure of dissimilarity (1 – TOM) was used to identify the clusters. The dissimilarity adjacency was taken to perform hierarchical average linkage clustering based on topological overlap and this was used to identify gene co-expression modules that grouped genes with similar expression pattern. The dynamic tree cut function was used to identify each gene modules with a minimum module size of 30 to preserve normality distribution assumption.

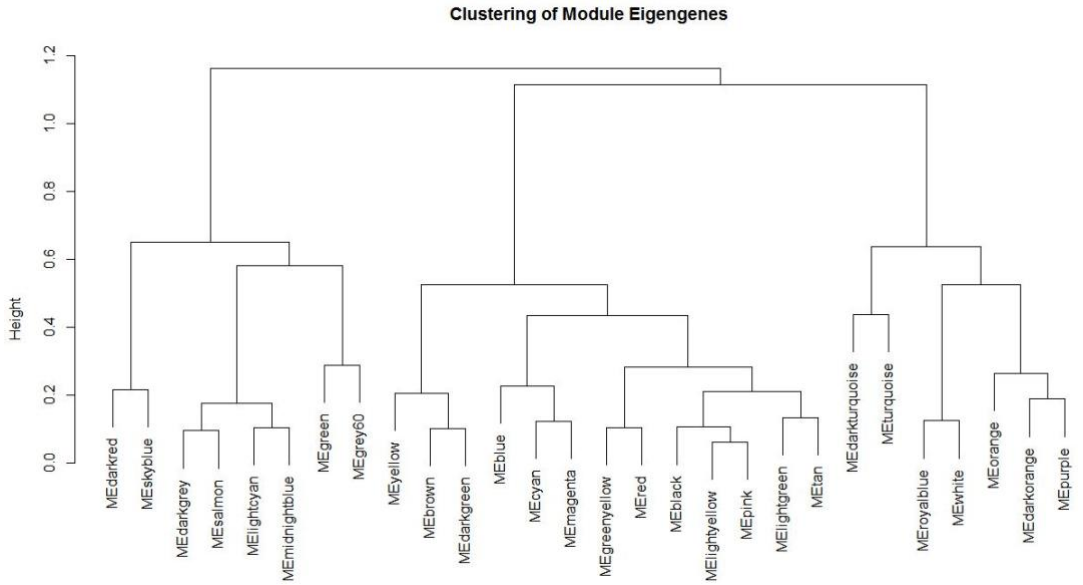
(Oldham, Horvath, & Geschwind, 2006) (Guo, Xiao, Guo, Dong, & Chen, 2017) (Liu, Jing, & Tu, 2016) The module tree diagram with eigengenes clustered, and a dendrogram with modules with a unique colour code assigned based on the module size were constructed. The same scale-free network construction approach above was used for all dataset from 5 drug treatment cases. The tree diagram and dendrogram were visualized for each drug treatment in figure 5.7 and the modules did not require any merging. Each vertical line in the dendrogram represents a single gene.

(a)

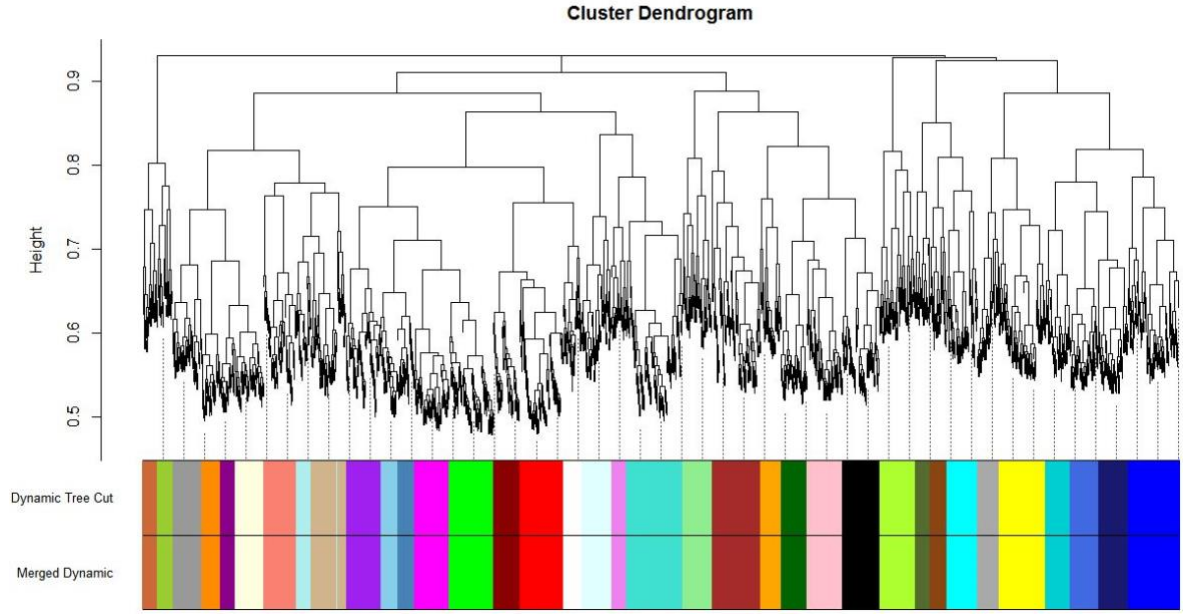


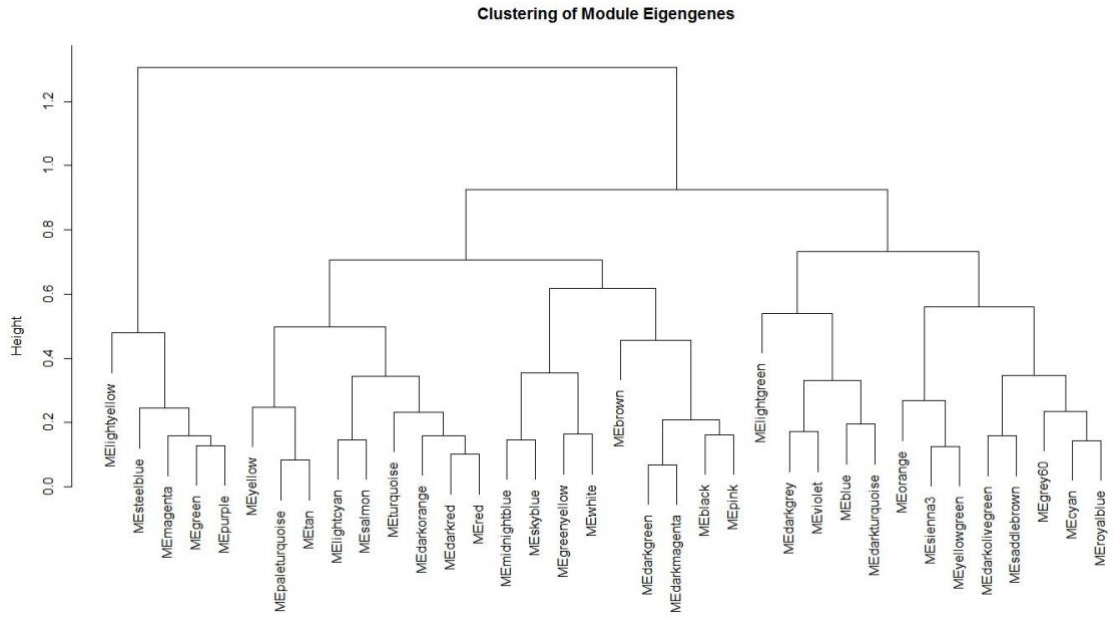
(b)


Cluster Dendrogram
Height
0.5
0.6
0.7
0.8
0.9
Dynamic Tree Cut
Merged Dynamic

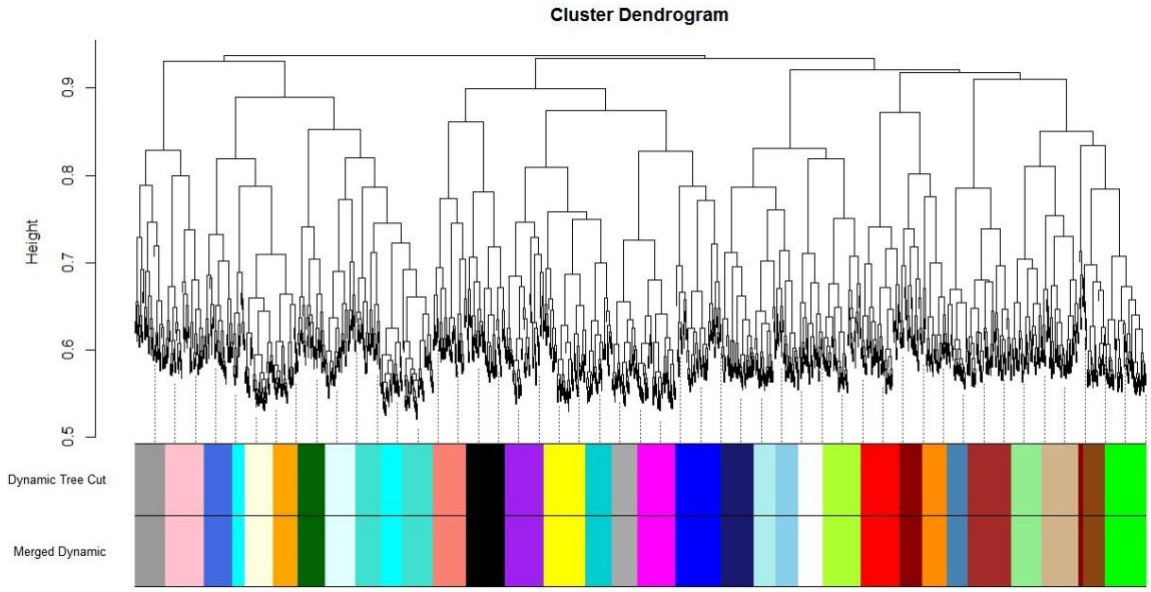

Clustering of Module Eigengenes
Height
0.0
0.2
0.4
0.6
0.8
1.0
1.2
MEdarkred
MEskyblue
MEdarkgrey
MEsalmon
MElightcyan
MEmidnightblue
MEgreen
MEgrey60
MEyellow
MEbrown
MEdarkgreen
MEblue
MEcyan
MEmagenta
MEgreenyellow
MEred
MEblack
MElightyellow
MEpink
MElightgreen
MEtan
MEdarkturquoise
MEturquoise
MEroyalblue
MEwhite
MEorange
MEdarkorange
MEpurple

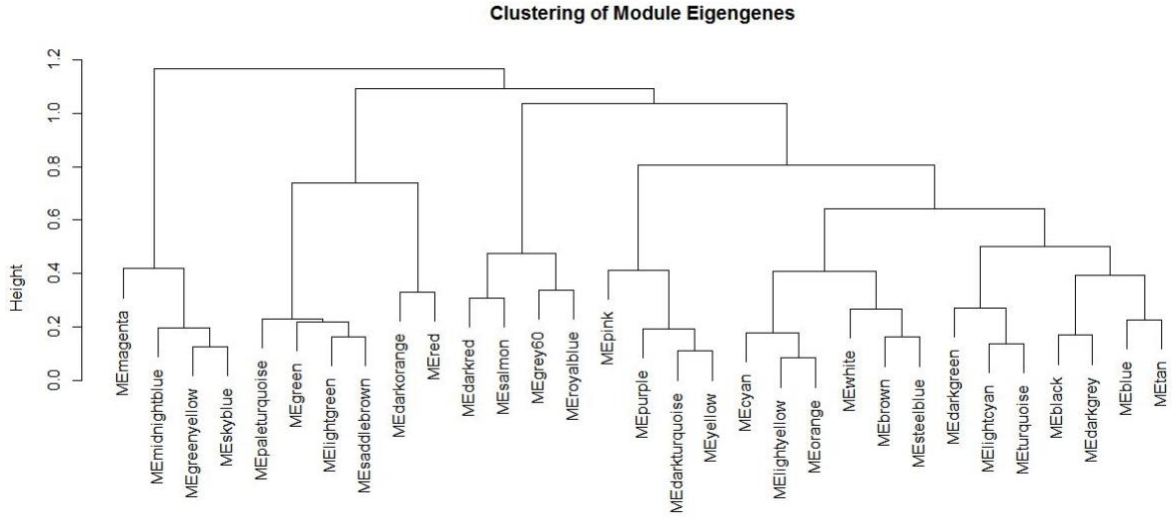
(c)


Cluster Dendrogram
Height
0.5
0.6
0.7
0.8
0.9
Dynamic Tree Cut
Merged Dynamic


Clustering of Module Eigengenes
Height
0.0
0.2
0.4
0.6
0.8
1.0
1.2
MElightyellow
MEsteelblue
MEmagenta
MEgreen
MEpurple
MEyellow
MEpaleturquoise
MEtan
MElightcyan
MEsalmon
MEturquoise
MEdarkorange
MEdarkred
MEred
MEmidnightblue
MEskyblue
MEgreenyellow
MEwhite
MEbrown
MEdarkgreen
MEdarkmagenta
MEblack
MEpink
MElightgreen
MEdarkgrey
MEviolet
MEblue
MEdarkturquoise
MEorange
MEsienna3
MEyellowgreen
MEdarkolivegreen
MEsaddlebrown
MEgrey60
MEcyan
MEroyalblue

(d)


Cluster Dendrogram
Height
0.5
0.6
0.7
0.8
0.9
Dynamic Tree Cut
Merged Dynamic


Clustering of Module Eigengenes
Height
0.0
0.2
0.4
0.6
0.8
1.0
1.2
MEmagenta
MEmidnightblue
MEgreenyellow
MEskyblue
MEpaleturquoise
MEgreen
MElightgreen
MEsaddlebrown
MEdarkorange
MEred
MEdarkred
MEsalmon
MEgrey60
MEroyalblue
MEpink
MEpurple
MEdarkturquoise
MEyellow
MEcyan
MElightyellow
MEorange
MEwhite
MEbrown
MEsteelblue
MEdarkgreen
MElightcyan
MEturquoise
MEblack
MEdarkgrey
MEblue
MEtan

(e)

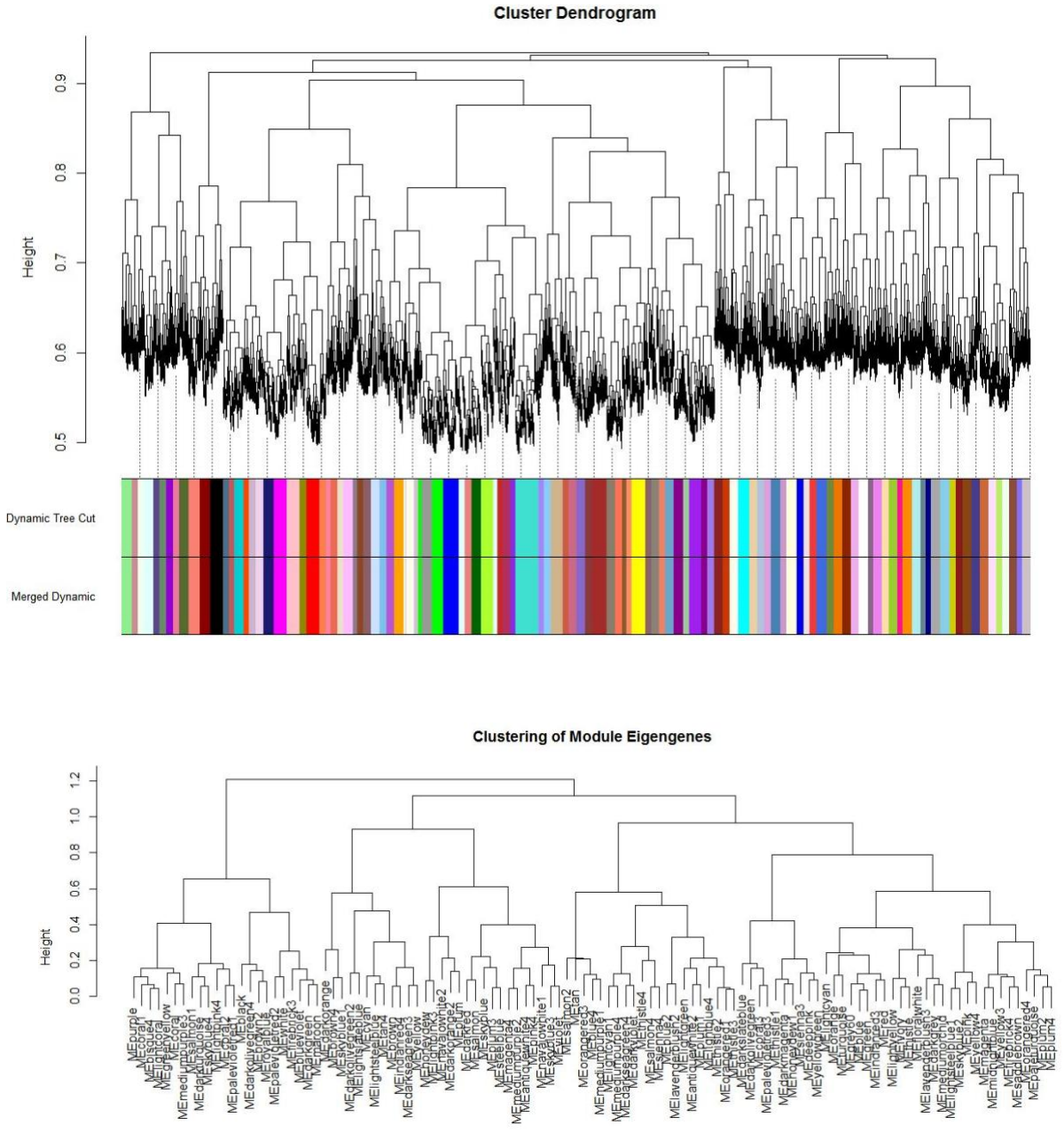


Figure 5.7 Identification of gene co-expression modules in MCF-7 after 5 different drug treatments via hierarchical average linkage clustering (Dynamic Tree Cut algorithm was used to identify modules. The identified modules are coded by unique colours. The dendrogram and tree diagram showing eigengenes clustering is also shown. (a) Fulvestrant treatment: 31 modules identified. (b) Monorden treatment: 28 modules identified. (c) Thioridazine treatment: 36 modules identified. (d) Trifluoperazine treatment: 31 modules identified. (e) Vorinostat treatment: 111 modules identified.

Using WGCNA package in R, the differentially expressed genes (DEGs) from 5 different drug treatments were analysed in terms of their co-expression pattern. The DEGs with similar expression pattern were assigned into modules via hierarchical average linkage clustering, and for each drug treatment, total number of modules (figure 5.7) were identified. The genes within each co-expressed module show high connectivity and correlation. Finally, the Entrez IDs of the genes within each module were retrieved for subsequent gene ontology analysis to associate each co-expressed module of genes to biological processes through enrichment test.

5.6 Biological interpretation

Once the Entrez ID of the genes were retrieved, functional enrichment analysis was performed for the selected drugs. The list of highly correlated co-expressed genes from 5.4.1 were retrieved from the modules generated in dendrogram, and they were related to functional annotation database such as Gene Ontology (GO). The purpose of GO functional annotation was to examine if the co-expressed genes from the modules were significantly enriched or impoverished in specific biological processes or functions in the GO terms. For this purpose, only the 3 cancer drugs namely, Fulvestrant, Monorden and Vorinostat were considered as their applications in breast cancer treatment were more well-known than the 2 non-cancer antipsychotic drugs. The co-expressed gene modules from the 3 cancer drugs were annotated in terms of its biological functions through GO enrichment analysis and the genes that were up and down regulated from figure 5.5 after treatment were mapped onto their respective co-expressed modules. Some co-expressed modules that represent specific biological function had down-regulated GT genes as its one of the genes co-expressed and some modules had up-regulated GT genes.

5.6.1 Gene ontology enrichment analysis

Functional annotation of the modules was done using “GO.db” annotation package which contains the latest version of gene ontologies database. The function, “GOenrichmentAnalysis” from the WGCNA package was utilized to perform GO enrichment analysis. The enrichment analysis function required Homo Sapiens organism database such as “org.hs.eg.db”, “AnnotationDbi” and “GO.db”. For each co-expression module, the various enrichment scores (p-value) was calculated

using Fisher exact test and the specified ontologies were returned accordingly. The GO enrichment result for the three cancer drugs were tabulated.

Table 5.3 GO enrichment significantly associated with WGCNA modules from Fulvestrant treatment. Only the modules with either up or down regulated glycosyltransferase genes co-expressed with other genes have been recorded. GO term is represented for each module.

Module	No. of genes	Term	p-value	GT
Blue	97	Regulatory region DNA binding (GO:0000975)	0.001068	Up
		Transcription regulatory region DNA binding (GO:0044212)	0.0011	
Brown	93	Positive regulation of neuron differentiation (GO:0045666)	0.005	Down
		Histone methyltransferase activity (GO:0042054)	0.005	
Dark red	51	Regulation of hormone biosynthetic process (GO:0043269)	0.0003	Up
		Regulation of ion transmembrane transport (GO:0034765)	0.0017	
Dark Turquoise	45	Mammary gland epithelial cell differentiation (GO:0060644)	0.005	Down
		Glucocorticoid receptor binding (GO:0035259)	0.005	
Green Yellow	69	Regulation of protein complex assembly (GO:0043254)	0.0013	Up
Light Green	53	Glycerophospholipid biosynthetic process (GO:0046474)	0.0032	Down
Orange	40	Oligodendrocyte development (GO:0014003)	0.009	Down
Purple	68	Positive regulation of transporter activity (GO:0032411)	0.006	Down
Saddle Brown	38	Regulation of hydrogen peroxide-induced cell death (GO:1903205)	0.003	Down
Tan	69	T-cell activation involved in immune response (GO:0002286)	0.003	Down
Yellow	91	Negative regulation of leukocyte migration (GO:0002686)	0.006	Up

Table 5.4 GO enrichment significantly associated with WGCNA modules from Monorden treatment. Only the modules with either up or down regulated glycosyltransferase genes co-expressed with other genes have been recorded. GO term is represented for each module.

Module	No. of genes	Term	p-value	GT
Black	80	Immunoglobulin mediated immune response (GO:0016064) Immunoglobulin production (GO:0002377)	0.0015 0.0021	Up
Blue	100	Wnt signalling pathway, planar cell polarity pathway (GO:0060071) Non-canonical Wnt signalling pathway (GO:0035567)	$8.78e^{-05}$ 0.00015	Down
Dark Grey	35	Positive regulation of neuron apoptotic process (GO:0043525)	0.0003	Down
Green	91	Signalling adaptor activity (GO:0035591)	0.004	Down
Green Yellow	70	Positive regulation of ATPase activity (GO:0032781)	0.008	Down
Midnight Blue	59	Glucose import (GO:0046323)	0.003	Down
Pink	77	Multicellular organism metabolic process (GO:0044236)	$1.71e^{-05}$	Up
Red	87	Hydrogen ion transmembrane transport (GO:1902600)	0.001	Down
Royal Blue	44	Activation-induced cell death of T cells (GO:0006924)	0.001	Down
Salmon	64	Nerve development (GO:0021675)	$5.17e^{-05}$	Down
Tan	64	Growth factor receptor binding (GO:0070851)	0.0004	Down
Turquoise	113	Regulation of lymphocyte activation (GO:0051249)	0.0027	Up

Table 5.5 GO enrichment significantly associated with WGCNA modules from Vorinostat treatment. Only the modules with either up or down regulated glycosyltransferase genes co-expressed with other genes have been recorded. GO term is represented for each module.

Module	No. of genes	Term	p-value	GT
Bisque4	54	Ribosome biogenesis (GO:0042254)	0.001	Down
Black	86	Regulation of cell division (GO:0051781)	0.0004	Down
Blue4	32	Glycoprotein biosynthetic process (GO:0010559)	0.003	Down
Brown	97	Nerve growth factor signalling pathway (GO:0038180)	0.003	Up
Dark Red	65	Motor neuron apoptosis (GO:0097049)	0.001	Up
Honeydew	42	RNA metabolic process regulation (GO:0051252)	0.005	Up
Ivory	57	Regulation of cell cycle (GO:0051726)	0.001	Up
Lavender Blush2	42	Activated T cell proliferation (GO:0050798)	0.005	Down
Light Coral	44	Positive regulation of small GTPase mediated signal transduction (GO:0051057)	0.006	Down
Light Cyan	70	Drug metabolic process (GO:0017144)	0.002	Down
Light Green	70	Regulation of histone phosphorylation (GO:0033127)	0.001	Down
Light Slate Blue	36	Polyamine biosynthetic process (GO:0006596)	0.001	Down
Light Yellow	68	Protein folding (GO:0006457)	0.003	Down
Maroon	50	DNA replication initiation (GO:0006270)	0.001	Down
Medium Orchid	48	Epithelial structure assembly (GO:0010669)	0.001	Down
Medium Purple1	36	Cellular response to wortmannin (GO:1904568)	0.006	Down
Medium Purple2	45	Positive regulation of cytokine production in immune system (GO:0002720)	0.003	Down
Orange	62	Regulation of protein transport (GO:0051223)	0.001	Down
Orange Red1	37	Glyceraldehyde-3-phosphate biosynthesis (GO:0046166)	0.006	Down
Orange Red4	58	B cell apoptotic process (GO:0001783)	0.004	Down
Pale Violet Red2	43	Protein localization (GO:0034613)	0.001	Up
Pale Violet Red3	51	Negative regulation of DNA binding (GO:0043392)	0.0004	Down

Plum2	54	Regulation of cAMP biosynthetic process (GO:0030817)	0.00001	Up
Plum3	44	Glucocorticoid catabolic process (GO:0006713)	0.007	Down
Plum4	32	Regulation of gluconeogenesis involved in cellular glucose homeostasis (GO:0090526)	0.005	Down
Royal Blue	69	Polyamine biosynthetic process (GO:0006595)	0.005	Down
Saddle Brown	61	Regulation of exocytosis (GO:0017157)	0.003	Up
Sienna3	60	Glycosphingolipid biosynthetic process (GO:0006688)	0.005	Up
Sky Blue	61	Cardiac right ventricle morphogenesis (GO:0003215)	0.004	Up
Tan	80	Sodium channel activity (GO:0005272)	0.001	Up
Thistle	43	Regulation of translation (GO:0006417)	0.006	Up
Thistle1	53	T-cell migration (GO:0072678)	0.002	Up
Thistle3	41	Cell surface pattern recognition receptor signalling pathway (GO:0002752)	0.006	Down
Violet	60	Transferase activity (GO:0016740)	0.002	Up
Yellow	97	Antigen processing and presentation of peptide antigen (GO:0048002)	0.002	Up
Yellow3	39	Regulation of TORC1 signalling (GO:1904263)	0.0001	Down
Yellow Green	59	Regulation of sodium ion transport (GO:0002028)	0.0002	Up

The significantly enriched functions for co-expressed modules with up regulated GT genes in Fulvestrant drug treatment (Table 5.3) were mostly regulatory associated biological processes. Some modules with down regulated GT genes were involved in activation and biosynthesis processes. Fulvestrant is known to treat hormone receptor (HR)-positive metastatic breast cancer and identifying GT genes that are responsible for breast cancer metastasis in the co-expressed modules generated by Fulvestrant treatment would present valuable insight into breast cancer treatment strategy. The up and down regulated GT genes in the co-expressed gene module produced by Fulvestrant can be identified and their mechanisms of action can be further studied.

Monorden drug treatment had a blue module with its co-expressed genes playing role in Wnt signalling pathways (Table 5.4) which is a pathway involved in carcinogenesis including breast and prostate cancers. The GT genes, MAN1B1 and ALG3 were involved in the blue module with Wnt signalling pathways and they were down regulated after treatment of Monorden. Turquoise module from Monorden treatment was significantly enriched in regulation of lymphocyte

activation with GCNT1 gene up regulated after treatment. Monorden is an inhibitor of heat shock protein 90 (HSP 90 stabilizes proteins involved in cancerous tumour growth) that has great potential as breast cancer treatment drug as they can suppress multiple oncogenic signalling pathways simultaneously as described in section 5.3.3.1. Indeed, one of the co-expressed modules was significantly enriched in Wnt signalling pathways and a few GT genes were down regulated in the co-expression module. Further research in the role of Monorden drug treatment in the specific module of co-expressed genes and GT genes within the module will be required to better understand how the GT genes and non-GT genes in the module co-express together to affect the breast cancer pathways.

Vorinostat drug treatment generated the most co-expressed modules and many enriched biological functions from GO annotation were returned (Table 5.5). The blue4 module from Vorinostat treatment was significantly enriched in glycoprotein biosynthesis process and the GT gene within the module has been down regulated after treatment. The orange module was enriched in regulation of protein transport and it had down regulated GT genes. Most of the biosynthesis processes that were enriched in Vorinostat treated co-expressed modules had down regulated GT genes, whereas activities such as sodium channel activity, T-cell migration, transferase and sodium ion transport had up regulated GT genes. Interestingly, thistle3 module was significantly enriched in cell surface pattern recognition of receptor signalling pathway and the GTs were down regulated after treatment.

After treatment with each of the three cancer drugs it has been shown here that they affect the biological functions associated with co-expression modules through alteration of the gene regulation in that module. The alteration of GT genes expression is of importance. These findings present useful insight into how the GT genes co-express with specific genes in a module to alter a biological function after treatment of specific cancer drugs. These findings could possibly be used in the future drug repurposing research and it is hoped that these results will aid in designing novel breast cancer therapeutic drugs.

5.7 References

1. Allemeersch, J., & Moreau, Y. (2004). Meta-analysis of microarrays.

2. Banerjee, S., Li, Y., Wang, Z., & Sarkar, F. H. (2008). Multi-targeted therapy of cancer by genistein. *Cancer letters, 269*(2), 226-242.
3. Clay, C. E., Monjazeb, A., Thorburn, J., Chilton, F. H., & High, K. P. (2002). 15-Deoxy-Δ12, 14-prostaglandin J2-induced apoptosis does not require PPARγ in breast cancer cells. *Journal of lipid research, 43*(11), 1818-1828.
4. Dimopoulos, M.-A., Mitsiades, C. S., Anderson, K. C., & Richardson, P. G. (2011). Tanespimycin as antitumor therapy. *Clinical Lymphoma, Myeloma and Leukemia, 11*(1), 17-22.
5. Drummond, D. C., Noble, C. O., Kirpotin, D. B., Guo, Z., Scott, G. K., & Benz, C. C. (2005). Clinical development of histone deacetylase inhibitors as anticancer agents. *Annu. Rev. Pharmacol. Toxicol., 45*, 495-528.
6. Guo, X., Xiao, H., Guo, S., Dong, L., & Chen, J. (2017). Identification of breast cancer mechanism based on weighted gene coexpression network analysis. *Cancer gene therapy*.
7. Hedberg, D. L., Houck, J. H., & GLUECK JR, B. C. (1971). Tranylcypromine-trifluoperazine combination in the treatment of schizophrenia. *American Journal of Psychiatry, 127*(9), 1141-1146.
8. Karve, S., Werner, M. E., Sukumar, R., Cummings, N. D., Copp, J. A., Wang, E. C., . . . Pacold, M. E. (2012). Revival of the abandoned therapeutic wortmannin by nanoparticle drug delivery. *Proceedings of the national academy of sciences, 109*(21), 8230-8235.
9. Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics, 9*(1), 559.
10. Langfelder, P., & Horvath, S. (2014). Tutorials for the WGCNA Package.
11. Lee, C. I., Goodwin, A., & Wilcken, N. (2017). Fulvestrant for hormone-sensitive metastatic breast cancer. *The Cochrane Library*.
12. Liu, J., Jing, L., & Tu, X. (2016). Weighted gene co-expression network analysis identifies specific modules and hub genes related to coronary artery disease. *BMC cardiovascular disorders, 16*(1), 54.
13. Luu, T. H., Morgan, R. J., Leong, L., Lim, D., McNamara, M., Portnow, J., . . . Gandara, D. R. (2008). A phase II trial of vorinostat (suberoylanilide hydroxamic acid) in metastatic breast cancer: a California Cancer Consortium study. *Clinical Cancer Research, 14*(21), 7138-7142.
14. Maira, S.-M., Stauffer, F., Schnell, C., & García-Echeverría, C. (2009). PI3K inhibitors for cancer treatment: where do we stand? : Portland Press Limited.
15. Munster, P., Thurn, K., Thomas, S., Raha, P., Lacevic, M., Miller, A., . . . Moasser, M. (2011). A phase II study of the histone deacetylase inhibitor vorinostat combined with tamoxifen for the

treatment of patients with hormone therapy-resistant breast cancer. *British journal of cancer, 104*(12), 1828.

16. Nathan, M. R., & Schmid, P. (2017). A Review of Fulvestrant in Breast Cancer. *Oncology and therapy, 5*(1), 17-29.
17. Oldham, M. C., Horvath, S., & Geschwind, D. H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the national academy of sciences, 103*(47), 17973-17978.
18. Overall, J. E., Hollister, L. E., Meyer, F., Kimbell, I., & Shelton, J. (1964). Imipramine and Thioridazine in Depressed and Schizophrenic Patients: Are there specific antidepressant drugs? *JAMA, 189*(8), 605-608.
19. Pacey, S., Wilson, R. H., Walton, M. I., Eatock, M., Hardcastle, A., Zetterlund, A., . . . Roels, B. (2011). A Phase I study of the Heat Shock Protein 90 inhibitor alvespimycin (17-DMAG) given intravenously to patients with advanced, solid tumors. *Clinical Cancer Research*, clincanres. 1927.2010.
20. Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S., & Smyth, G. K. (2016). Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *The annals of applied statistics, 10*(2), 946.
21. Picco, G., Julien, S., Brockhausen, I., Beatson, R., Antonopoulos, A., Haslam, S., . . . Taylor-Papadimitriou, J. (2010). Over-expression of ST3Gal-I promotes mammary tumorigenesis. *Glycobiology, 20*(10), 1241-1250.
22. Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature genetics, 32*, 496.
23. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research, 43*(7), e47-e47.
24. Sánchez, A., & de Villa, M. (2008). A tutorial review of microarray data analysis. *Universitat de Barcelona*.
25. Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology, 3*(1), 1-25.
26. Smyth, G. K. (2005). Limma: linear models for microarray data *Bioinformatics and computational biology solutions using R and Bioconductor* (pp. 397-420): Springer.

27. Smyth, G. K., Ritchie, M., Thorne, N., & Wettenhall, J. (2005). LIMMA: linear models for microarray data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Statistics for Biology and Health.
28. Strobl, J. S., & Peterson, V. A. (1992). Tamoxifen-resistant human breast cancer cell growth: inhibition by thioridazine, pimozide and the calmodulin antagonist, W-13. *Journal of Pharmacology and Experimental Therapeutics*, *263*(1), 186-193.
29. Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D., & Altman, R. B. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, *18*(11), 1454-1461.
30. van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2017). Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*, bbw139.
31. Wang, D., & DuBois, R. N. (2006). Prostaglandins and cancer. *Gut, 55*(1), 115-122.
32. Warrell Jr, R. P., Frankel, S. R., Miller Jr, W. H., Scheinberg, D. A., Itri, L. M., Hittelman, W. N., . . . Jakubowski, A. (1991). Differentiation therapy of acute promyelocytic leukemia with tretinoin (all-trans-retinoic acid). *New England Journal of Medicine, 324*(20), 1385-1393.
33. Wei, J.-W., Hickie, R. A., & Klaassen, D. J. (1983). Inhibition of human breast cancer colony formation by anticalmodulin agents: trifluoperazine, W-7, and W-13. *Cancer chemotherapy and pharmacology*, *11*(2), 86-90.
34. Yun, J., Lv, Y., Yao, Q., Wang, L., Li, Y., & Yi, J. (2012). Wortmannin inhibits proliferation and induces apoptosis of MCF-7 breast cancer cells. *European journal of gynaecological oncology, 33*(4), 367-369.
35. Zagouri, F., Sergentanis, T. N., Chrysikos, D., Papadimitriou, C. A., Dimopoulos, M.-A., & Psaltopoulou, T. (2013). Hsp90 inhibitors in breast cancer: a systematic review. *The Breast, 22*(5), 569-578.

6 CONCLUSIONS AND FUTURE WORK

Gene expression technology combined with bioinformatics applications can be used to study the effect of different therapeutic drugs on breast cancer cells. The examination of gene expression patterns after various drug treatments provides a better understanding of how a specific drug can affect the regulation of critical genes that play significant roles in breast cancer proliferation.

The first task of the present work involved obtaining raw gene expression data from the CMap database. The expression data used were generated from microarray technology and consisted of MCF-7 breast cancer cell samples treated with 29 different therapeutic drugs, including anticancer and non-cancer drugs. Several pre-processing steps were followed, including data cleaning, quality checking, normalization, and batch effect removal, to obtain a gene expression matrix for differential expression and co-expression analysis.

Differential gene expression analysis was performed based on the reliable gene expression matrix obtained previously, and drug-treated samples with DEGs were filtered using several statistical protocols. To perform a more robust analysis, a two-group Wilcoxon rank sum meta-analysis was performed based on the previously selected samples to identify drug candidates that had a significant effect on the regulation of GT genes. Three anticancer drugs, including Fulvestrant, Monorden, Vorinostat, and the two antipsychotic drugs trifluoperazine and thioridazine, passed the rank sum test. A set of GT genes was identified, and their expression patterns were visualized as heatmaps.

Finally, the WGCNA method was applied to the five drugs selected to identify co-expressed gene modules. Various co-expressed modules were identified for each drug and the three anticancer drugs were then selected for subsequent GO enrichment analysis to correlate co-expressed modules with biological functions. The co-expression modules were selected based on the presence of up- or down-regulated GT genes in the module. For each drug treatment, a number of different modules with associated biological functions were identified. Fulvestrant treatment generated co-expressed modules that were involved in biosynthesis, transportation regulation, and apoptosis. Monorden treatment generated co-expressed modules that were involved in the Wnt signalling pathway, which is an oncogenic cancer signalling pathway. Interestingly, some of these

GT genes were down-regulated, suggesting a possible interaction between GT and other genes in this co-expression module, acting to suppress this cancer signalling pathway. Monorden could therefore be effective in breast cancer treatment. Vorinostat treatment generated co-expression modules that were involved in glycoprotein biosynthesis, protein transportation, and cell surface pattern recognition for receptor signalling pathways. The GT genes involved in these biological functions were down-regulated after Vorinostat treatment.

The most important limitation of this research is the insufficient number of biological replicates to enhance the statistical significance of analyses and generate more reliable and robust results. For a typical gene expression analysis, not enough replicates were used in this study. To overcome this challenge to some extent, an additional meta-analysis incorporating a two-group Wilcoxon rank sum test was performed. The usage of Wilcoxon rank sum test removed the bias of selecting small number of replicates in this study. However, in the future, incorporation of a greater number of biological replicates would doubtless improve the study and generate more reliable results.

The methods used here to identify specific DEGs and DCMs along with their enriched biological functions will provide useful insight into future drug repurposing strategies to develop novel therapeutic drug candidates that will target specific genes and pathways in human breast cancer. In terms of future work, further study of DEGs involved in co-expressed modules and their annotated biological functions would be needed, and an in-depth pathway analysis of the genes and co-expressed modules identified may contribute to the greater understanding of breast cancer mechanisms and the development of novel breast cancer drugs.

APPENDICES

Appendix A: Copyrights materials

Figure 1.4 - * Permission to reproduce this figure has been granted by the author of "In silico gene expression profiling in Cannabis sativa". Creative Commons Attribution License permits unrestricted reproduction of the contents provided the original work is properly cited [accessed on 23rd February 2018]. Reprinted from the paper's copyright section: "Copyright: © 2017 Massimino L. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited."

Figure 2.2 - * Permission to reproduce this figure has been granted by the author of "Analysis of microarray experiments of gene expression profiling" [accessed on 28th February 2018]. Reprinted from the paper's copyright section: "Articles published in these journals are in the public domain and may be used and reproduced without special permission." Available at: https://www.ncbi.nlm.nih.gov/pmc/about/copyright/.

Figure 2.4 - * Permission to reproduce this figure has been granted by the author of "A tutorial review of microarray data analysis" [accessed on 16th February 2018]. Reprinted from the copyrights information of University of Barcelona about this material: "We do not need to ask for permission in these cases: **materials with free licenses**, generally, you can use any material that is disseminated with a free license, at least for no commercial purpose, but you always have to recognize the author and show the notice of the legal situation." Available at: http://crai.ub.edu/ca/que-ofereix-el-crai/drets-d-autor-i-propietat-intellectual-i-acces-obert/us-recursos-informacio-aliens.

Figure 2.6 - * Permission to reproduce this figure has been granted by the author of "WGCNA: An R package for weighted correlation network analysis" [accessed on 20th February 2018]. Reprinted from the paper's copyright section: "© Langfelder and Horvath; licensee BioMed Central Ltd. 2008. This article is published under license to BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited."

Figure 2.7 - * Permission to reproduce this figure has been granted by the author of "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists" [accessed on 3rd March 2018]. Reprinted from the paper's copyright section: "Copyright © 2008 The Author(s). This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited."

Figure 3.1 - * Permission to reproduce this figure has been granted by the author of "Essentials of Glycobiology" [accessed on 17th March 2018]. Reprinted from the copyright information: "Some of the content found in Bookshelf is authored and published by the National Center for Biotechnology Information (NCBI) or other institution of the U.S. government. No permission is needed to reproduce or distribute this type of content, but the authoring institute or agency must be given appropriate attribution."

Figure 3.2 - * Permission to reproduce this figure has been granted by the author of "Glycan profiling of adult T-cell leukemia (ATL) cells with the high-resolution lectin microarrays." [accessed on 17th March 2018]. Reprinted from the paper's copyright section: "© 2013 Iha and Yamada; licensee InTech. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits

unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited."

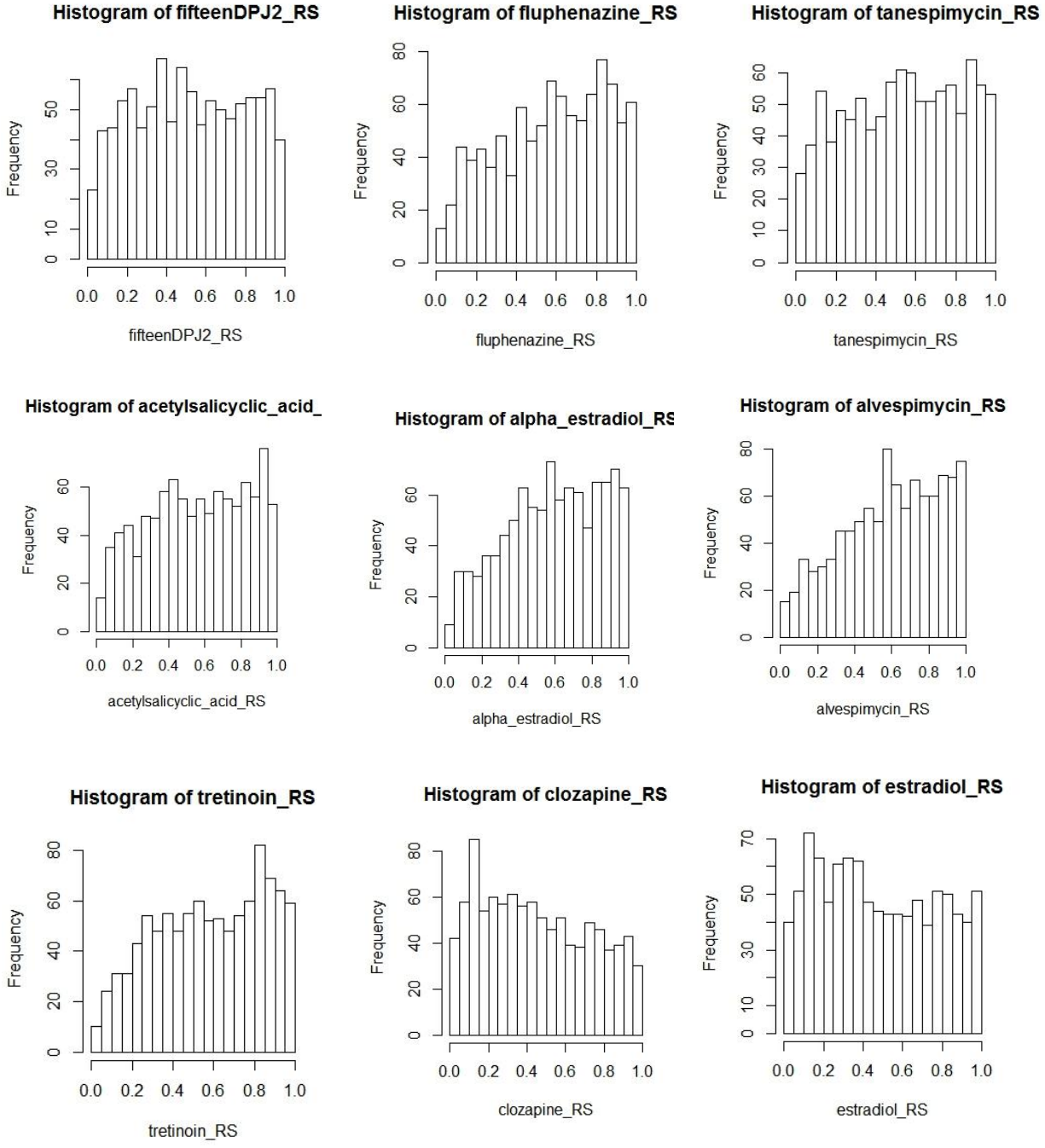
Figure 3.3 - * Permission to reproduce this figure has been granted by the author of "Essentials of Glycobiology" [accessed on 17th March 2018]. Reprinted from the copyright information: "Some of the content found in Bookshelf is authored and published by the National Center for Biotechnology Information (NCBI) or other institution of the U.S. government. No permission is needed to reproduce or distribute this type of content, but the authoring institute or agency must be given appropriate attribution."

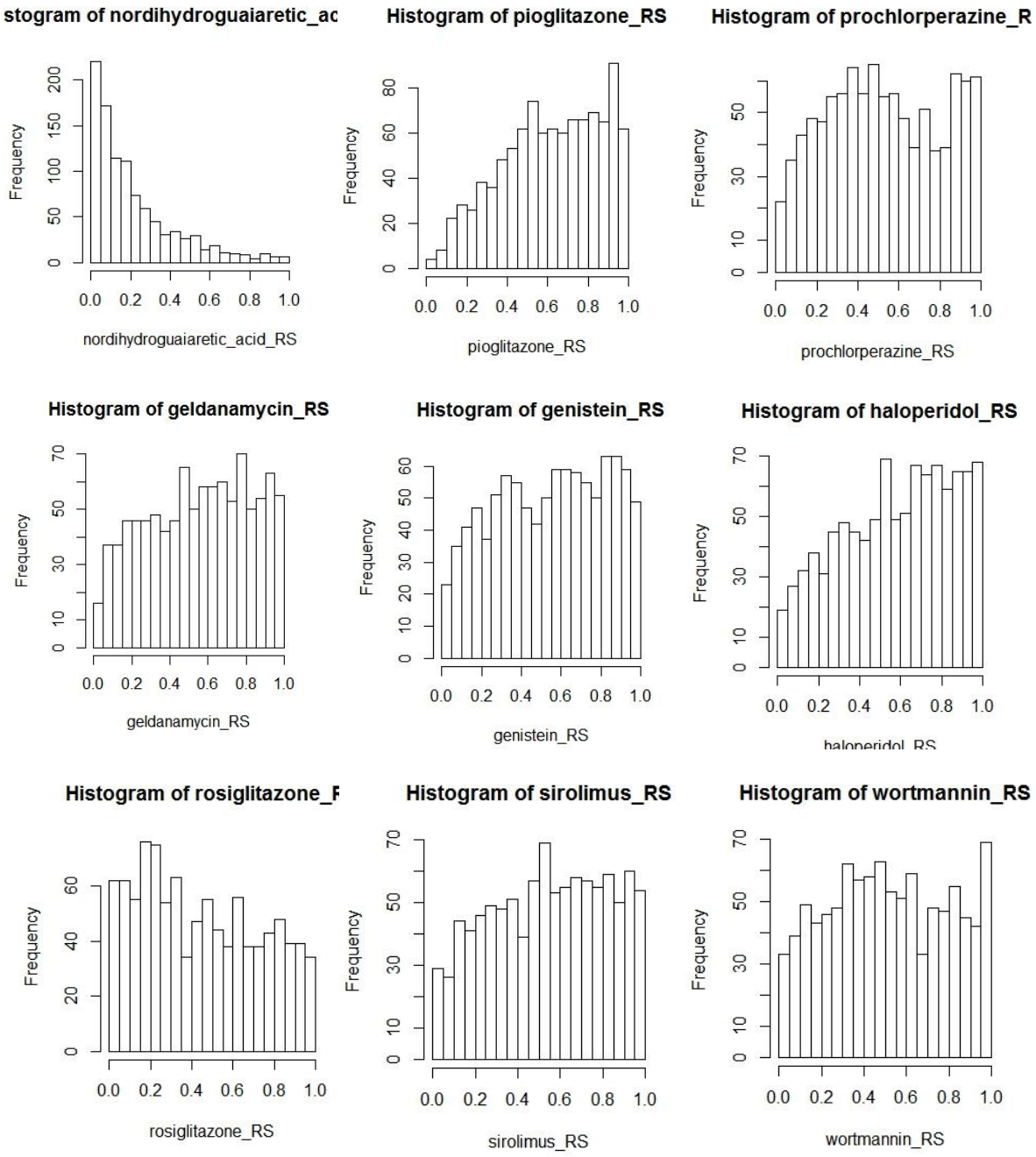
Figure 3.4 - * Permission to reproduce this figure has been granted by the author of "Glycan gene expression signatures in normal and malignant breast tissue; possible role in diagnosis and progression" [accessed on 26th March 2018]. Reprinted from the paper's copyright information: "All Molecular Oncology articles are published under the terms of the Creative Commons Attribution License (CC BY) which allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC BY license permits commercial and non-commercial re-use of an open access article, as long as the author is properly attributed. Authors also grant any third party the right to use the article freely as long as its original authors, citation details and publisher are identified."

Available at: https://febs.onlinelibrary.wiley.com/hub/journal/18780261/about/permissions

Appendix B: Additional charts and tables

Figure 1: The remainder p-value distribution histograms obtained from the meta-analysis (Rank Sum Test) for each drug treatment from figure 5.3.




stogram of nordihydroguaiaretic_ac
Frequency
nordihydroguaiaretic_acid_RS
Histogram of pioglitazone_RS
Frequency
pioglitazone_RS
Histogram of prochlorperazine_R
Frequency
prochlorperazine_RS
Histogram of geldanamycin_RS
Frequency
geldanamycin_RS
Histogram of genistein_RS
Frequency
genistein_RS
Histogram of haloperidol_RS
Frequency
haloperidol_RS
Histogram of rosiglitazone_F
Frequency
rosiglitazone_RS
Histogram of sirolimus_RS
Frequency
sirolimus_RS
Histogram of wortmannin_RS
Frequency
wortmannin_RS

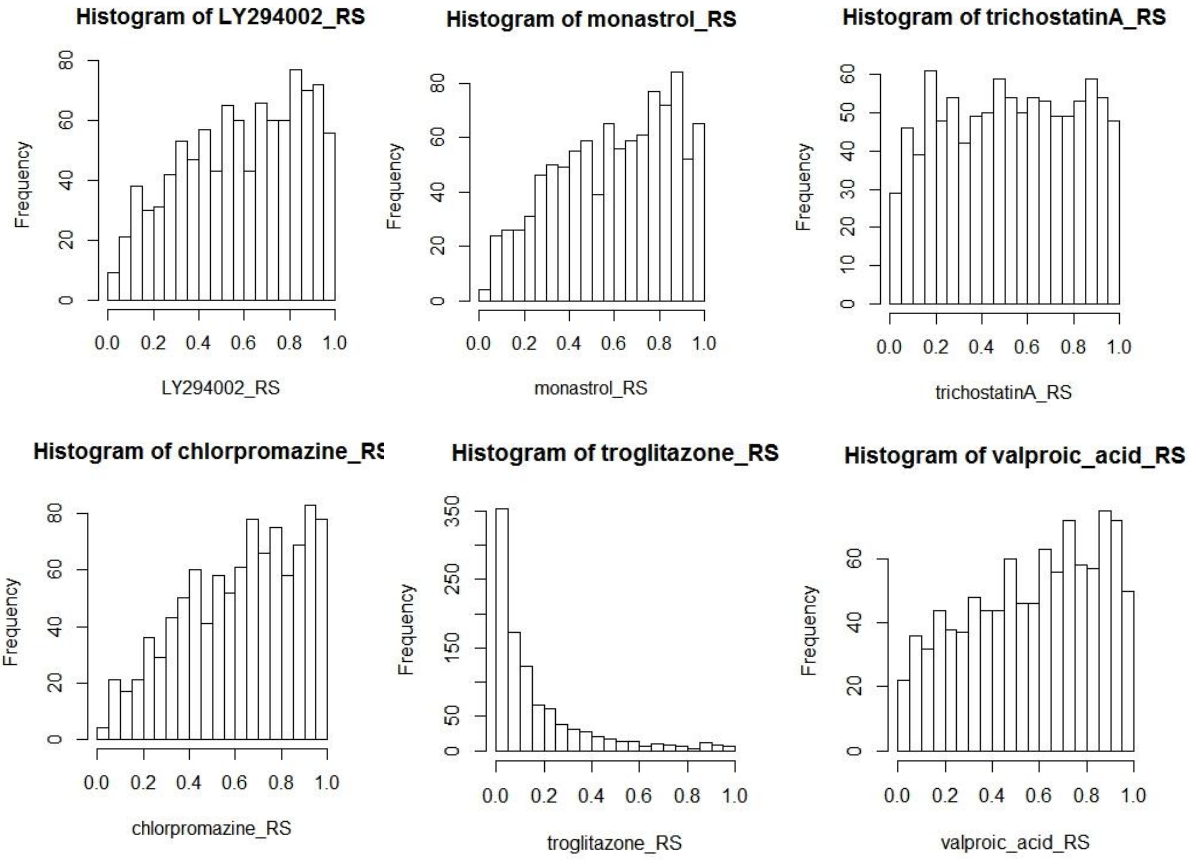

Histogram of LY294002_RS
Frequency
LY294002_RS
Histogram of monastrol_RS
Frequency
monastrol_RS
Histogram of trichostatinA_RS
Frequency
trichostatinA_RS
Histogram of chlorpromazine_RS
Frequency
chlorpromazine_RS
Histogram of troglitazone_RS
Frequency
troglitazone_RS
Histogram of valproic_acid_RS
Frequency
valproic_acid_RS

Table 1: A list of 29 drugs that have passed the selection criteria.

Drug Name	Indication	Target Protein	Chemical Formula with PubChem CID
Trichostatin A	Cancer	Histone Deacetylase (inhibitor)	$C_{17}H_{12}N_2O_3$ / 444732
Vorinostat	Cutaneous T-Cell Lymphoma	Histone Deacetylase (inhibitor)	$C_{14}H_{20}N_2O_3$/ 5311
Valproic Acid	Anticonvulsant	Histone Deacetylase, Calcium Channel	$C_8H_{16}O_2$/ 3121
LY-294002	Cancer	Phosphoinositide-3-Kinase (inhibitor)	$C_{19}H_{17}NO_3$/ 3973
Wortmannin	Hematologic Malignancies	Phosphatidylinositol-3-Kinase, Delta Isoform (inhibitor)	$C_{23}H_{24}O_8$/ 312145
Sirolimus	Organ Rejection	Threonine-Protein Kinase mTOR (inhibitor)	$C_{51}H_{79}NO_{13}$/ 5284616
Fulvestrant	Breast Cancer	Estrogen Receptor (antagonist)	$C_{32}H_{47}F_5O_3S$/ 104741
Estradiol	Urogenital Symptoms	Estrogen Receptor (antagonist)	C18H24O2/ 5757
Haloperidol	Antipsychotic	Dopamine Receptor (antagonist)	$C_{21}H_{23}CIFNO_2$/ 3559
Prochlorperazine	Psychotic Disorders	Dopamine Receptor (antagonist)	$C_{20}H_{24}CIN_3S$/ 4917

Clozapine	Treatment-Resistant Schizophrenia	5-HT Receptor (antagonist)	$C_{18}H_{19}CIN_4$/ 2818
Chlorpromazine	Schizophrenia, Nausea and Vomiting	5-HT2 Receptor (antagonist)	$C_{17}H_{19}CIN_2S$/ 2726
Fluphenazine	Psychotic Disorders	Dopamine Receptor (antagonist)	$C_{22}H_{26}F_3N_3OS$/ 3372
Trifluoperazine	Anxiety Disorders	Dopamine Receptor (antagonist)	$C_{21}H_{24}F_3N_3S$/ 5566
Thioridazine	Schizophrenia and Generalized Anxiety Disorders	Dopamine Receptor (antagonist)	$C_{21}H_{26}N_2S_2$/ 5452
Tanespimycin	Breast Cancer, Melanoma, Multiple Myeloma	Heat Shock Protein (HSP90 inhibitor)	$C_{31}H_{43}N_3O_8$/ 6440175
Geldanamycin	Oncogenesis, Angiogenesis, Apoptosis	Heat Shock Protein (HSP90 inhibitor)	$C_{29}H_{40}N_2O_9$/ 5288382
Alvespimycin	Ovarian Cancer, Breast Cancer, Leukemia	Heat Shock Protein (HSP90 inhibitor)	$C_{32}H_{48}N_4O_8$/ 5288674
Monorden	Oncogenesis, Angiogenesis, Apoptosis	Heat Shock Protein (HSP90 inhibitor)	$C_{18}H_{17}CIO_6$/ 6323491
Alpha-estradiol	Hair Loss	Androgen Receptor (antagonist)	$C_{18}H_{24}O_2$/ 68570
Troglitazone	Type II Diabetes Mellitus	Peroxisome Proliferation Activated Receptor	$C_{24}H_{27}NO_5S$/ 5591

Rosiglitazone	Type II Diabetes Mellitus	Peroxisome Proliferation Activated Receptor	$C_{18}H_{19}N_3O_3S$/ 77999
Pioglitazone	Type II Diabetes Mellitus	Peroxisome Proliferation Activated Receptor	$C_{19}H_{20}N_2O_3S$/ 4829
Nordihydroguaiaretic Acid	Antioxidant	Acetyl-CoA Acetyltransferase, Lipoxygenase (inhibitor)	$C_{18}H_{22}O_4$/ 4534
15-Delta Prostaglandin J2	Inducer of Gene Expression	-	$C_{20}H_{28}O_3$/ 5311211
Tretinoin	Acute Promyelocytic Leukemia	Retinoic Acid Receptor (antagonist)	$C_{20}H_{28}O_2$/ 444795
Acetylsalicylic Acid (Aspirin)	Various Forms of Pain, Inflammation	Cyclo-Oxygenase-1 (inhibitor)	$C_9H_8O_4$/ 2244
Monastrol	Spindle Bipolarity	Kinesin-5 (KIF11, Kinesin EG5 inhibitor)	$C_{14}H_{16}N_2O_3S$/ 2987927
Genistein	Prostate Cancer	DNA Topoisomerase 2-alpha inhibitor, Estrogen Receptor	$C_{15}H_{10}O_5$/ 5280961